XLIM-MS

Towards the Development of a Novel approach to Cross-linking Mass Spectrometry

Juliette M.B. James

University College London

Institute of Structural and Molecular Biology

Wellcome Trust 4 year Interdisciplinary PhD Programme in Structural, Computational and Chemical Biology

Supported by wellcome trust

For those who changed my world and gave me purpose.

For Michael & Asha.

Contents

Declaration

I, Juliette James, declare that the work presented in this thesis to be my own. Where information has been derived from other sources I confirm that it has been explicitly referenced within the text.

List of Figures

List of Tables

List of Abbreviations

BS3	BisSulfoSuccinimdylSuberate
BSA	Bovine Serum Albumin
BSG	Bis-sulfosuccinimidyl glutarate cross-linker
CDI	1,1'-carbonyldiimidazole cross-linker
CID	Collision Induced Dissociation
CCS	Collision Cross Section
CDF	Cumulative Probability Function
DDA	Data Dependent Acquisition
DSSO	Disuccinimdyl sulfoxide cross-linker
DMSO	Dimethylsulfoxide
DSS	DiSuccinimdylSuberate
EM	Electron Microscopy
ESI	ElectroSpray Ionisation
FWHM	Full Width Half Maximum height
FT-ICR	Fourier Transform Ion Cyclotron Resonance
HCD	High energy Collisional Dissociation
HDC	High Duty Cycle
HD-DDA	High Definition Data Dependent Acquisition
HSA	Human Serum Albumin
IM	Ion Mobility
IM-DDA	Ion Mobility Data Dependent Acquistion
IM-MS	Ion Mobility Mass Spectrometry
KDE	Kernel Density Estimation

LC-MS/MS	Liquid Chromatography Tandem Mass Spectrometry
LDA	Linear Discriminant Analysis
LIT	Linear Ion Trap
MALDI	Matrix Assisted Laser Desorption/Ionisation
MGF	Mascot Generic Format
m/z	Mass to Charge ratio
NCE	Normalised Collision Energy
NHS	N-HydroxySuccinimide
NMR	Nuclear Magnetic Resonance
oa-ToF	Orthogonal Acceleration Time of Flight
QToF	Quadrupole Time of Flight
PDF	Probability Density Function
PPM	Parts per Million
RMSD	Root Mean Squared Deviation
RNAPII	RNA Polymerase II
SASD	Solvent Accessible Surface Distance
SCX	Strong Cation Exchange Chromatography
SEC	Size Exclusion Chromatography
SNR	Signal to Noise Ratio
sulfo-SDA	Sulfosuccinimidyl 4,4-azipentanoate cross-linker
TIC	Total Ion Current
ToF	Time of Flight
TWIG	Travelling Wave Ion Guide
TWIM	Travelling Wave Ion Mobility
ToF	Time of Flight
UPLC	Ultra-high Performance Liquid Chromatography
XLMS	Crosslinking Mass Spectrometry

Abstract

In the cellular environment proteins form diverse and complex networks of interactions that control a variety of vital functions. Understanding the structure of these multi-subunit assemblies is key to understanding their function. To date the majority of structural information has been obtained through crystallographic studies, electron microscopy and NMR. However, the size and dynamics of these macro-molecular machines often precludes their analysis by such traditional methods. Cross-linking mass spectrometry offers a complementary structural technique. The distance restraints provided by the technique are formed in solution and offer more native structural information. Mapping of these restraints allows determination of the relative positions of amino acid residues in wider three dimensional structures.

To date the analysis of cross-linked peptides has almost exclusively been conducted with Orbitrap analysers. As a result most of the software applications designed to identified cross-link:spectrum matches have been developed with data from this type of analyser. Here we present an optimised protocol for the analysis of cross-linked samples using a Quadrupole Time-of-Flight mass spectrometer (QToF). We show that existing software can be configured to analyse QToF data with minimal adaptations. We evaluate the usefulness of the xQuest linear discriminant score in determining genuine cross-link:spectrum assignments.

The increased size and charge of crosslinked peptides compared to their un-crosslinked counterparts makes them ideal candidates for separation by ion mobility mass spectrometry. We take advantage of the unique geometry of the Triwave Stacked Ring Ion Guide to explore the effects of ion mobility separation on both precursors and fragment ions from cross-linked samples. To evaluate the sequence coverage and signal to noise ratio of identified cross-link:spectrum matches we present two computational solutions: ValidateXL and AnnotateXL.

Impact Statement

The RCSB Protein Data Bank currently stores over 140,000 structures. 16% of these contain over 1000 amino acids. One of the largest structures, the human nuclear pore complex has a combined total of over 19,000 residues. The model, published in 2015, used a combination of structural methods to gain insight into the structure of this 110 MDa macro-molecular machine.[2] In order to study larger and more dynamic assemblies structural biologists require a diverse range of experimental methods.

Cross-linking mass spectrometry provides a complementary addition to the structural biologists tool kit. The method may be applied independently of protein size or dynamics. Consequently, it is a complimentary technique to more traditional structural approaches such as NMR and Xray crystallography. To date almost all cross-linking mass spectrometry is conducted using Orbitrap analysers. In this work we have extended the field of cross-linking mass spectrometry to include the use of QToF mass spectrometers. In addition to a protocol for cross-linking on a QToF we have also shown that current cross-linking software, designed for use with Orbitrap style data, can be easily adapted for use with QToF data. This work has been submitted for publication and will contribute to the knowledge and refinement of this experimental technique.

To further enhance cross-linking analysis using a QToF geometry we have also conducted a study of the effects of ion mobility separation on cross-linked protein digests. This provides the ground work for future scholarships to continue the evaluation of the method, which may in turn lead to improved identification rates of cross-links from complex mixtures. By improving the discovery rate, more in-depth structural information can be obtained for the protein assembly being studied.

In addition to the experimental developments two computational solutions for cross-link

validation have also been created. These projects are available to download from `https://www.github.com/ThalassinosLab` and are offered under a GNU General Public License v3.0. Users of the software are permitted to use, modify and distribute changes to the software freely but without guarantee or warranty. The first piece of software, ValidateXL, offers xQuest users an extra layer of quality control to assess the validity of cross-link identifications. It reduces the time scale of an experiment by focusing manual validation to where it is needed most. The second offering, AnnotateXL, annotates any MS/MS spectrum of a cross-linked peptide. It can be used with data generated from any mass spectrometer and can be adapted for all cross-links.

These methods make cross-linking analysis accessible to a broader base of mass spectrometry laboratories both inside and outside academia. By enabling a greater number of mass spectrometry laboratories to use cross-linking as a structural biology tool a wider range of protein structures will be solved. This work is of benefit to anyone wishing to gain insight into protein structure using cross-linking mass spectrometry.

Chapter 1

Introduction

1.1 History of Mass Spectrometry

The first separation and measurement of charged ions was documented in 1912 by J.J. Thomson and F.W. Aston.[110] By deflecting a positively charged beam of particles, known as a positive ray, on to a photographic plate Thompson and Aston recorded the first mass spectra of Neon-20 and Neon-22. This pioneering evidence for the existence of isotopes from a stable element was met with significant scepticism as these compounds could not be separated by distillation.[112] Despite these early deliberations, mass spectrometry continued to evolve predominantly in the field of physical chemistry for the next sixty years.

The first biological molecules became accessible to mass spectrometry following the advent of Fast Atom Bombardment in 1981.[4] A solution containing the sample of interest was bombarded by a beam of high energy neutral atoms. The ejected ion beam was then accelerated into the instrument for analysis. This technique was used to generate the first deprotonated and protonated mass spectra for Bovine insulin.[3]

The continued development of mass spectrometry has resulted in the award of a number of Nobel prizes. The first of these was the Prize for Physics, awarded to J.J. Thomson in 1906 for the discovery of the electron. In addition, his work with positive rays enabled the discovery of the first isotope. In 1922 the Nobel Prize for Chemistry was awarded to F.W. Aston for his continuation of the work in non-radioactive elements, leading to the formulation of the "whole number rule". In the development of instrumentation the 1989 Prize for Physics

was shared with W. Paul and H.G. Dehmelt who received half of the award for their work on trapping ions.

Amongst the most notable for the field of biological mass spectrometry was the 2002 Nobel Prize in Chemistry. This was awarded for "the development of methods for identification and structural analyses of biological macromolecules". The Prize was shared between J.B. Fenn and K. Tanaka for their work on ionisation techniques. Fenn's discovery of Electrospray Ionisation (ESI) enabled the analysis of larger protein ions.[27] The technique combines the separation of individual protein molecules with the accumulation of charge to allow them to be introduced into the mass spectrometer.

1.2 Electrospray Ionisation

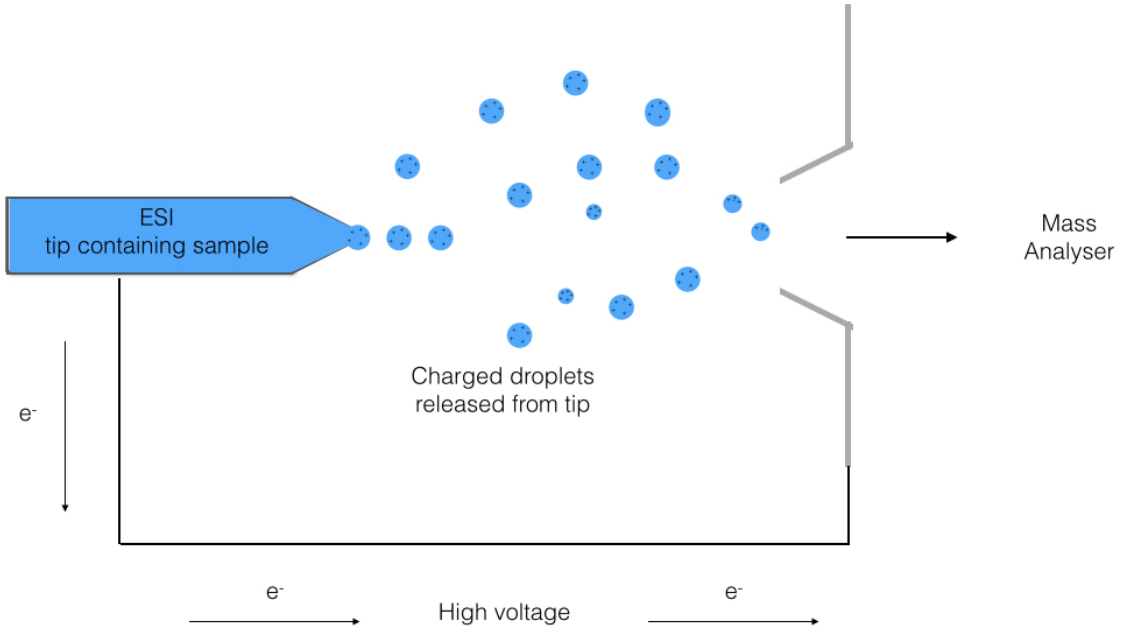


Figure 1.1: Schematic representation of electrospray ionisation. High voltage current is applied to the end of the spray tip producing charged droplets of volatile solvent. The solvent evaporates leaving behind charged particles that are deflected by electromagnetic fields through the mass spectrometer.

Electrospray ionisation (ESI) is a soft ionisation technique that generates a continuous beam of ions. The sample is placed in a volatile buffer that passes through a capillary tube under

atmospheric pressure. A strong electromagnetic field is applied to the end of the tube causing the accumulation of charge on the liquid at the end of the capillary. As the charge accumulates the Coloumbic attraction overcomes the surface tension of the solvent and the droplet deforms forming a Taylor cone which is expelled from the tip of the capillary.

Multiply charged ions are generated from the droplets as they pass through a curtain of heated inert gas (Figure 1.1). As the solvent molecules evaporate the droplets approach the theoretical maximum amount of charge that can be held by a drop of liquid. This is known as the Rayleigh limit.[26] Beyond this limit the droplets dissociate to form a beam of gas phase ions that enter the mass analyser. The ions are formed by protonation ($M + zH^z$) or deprotonation ($M - zH^z$), where M is mass and z is charge and H represents a proton. For macromolecules such as proteins this results in a characteristic peak distribution in a spectrum known as the "charge state envelope", in which adjacent peaks represent different charge states of the same ion.

The measured mass is a mass to charge ratio (m/z). In order to calculate the actual mass, the charge state of the ion must be known. Adjacent peaks in the charge state envelope differ by $z = 1$. Hence the mass can be calculated using the following set of equations:

$$m_1 = \frac{(M + z)}{z} \tag{1.1}$$

$$m_2 = \frac{(M + z + 1)}{z + 1} \tag{1.2}$$

Where m_1 and m_2 are the experimentally determined m/z ratio of two neighbouring peaks and m_2 is the peak with the lowerm/z. The mass of m_1 is the mass plus the charge (z) divided by the charge. The mass of m_2 may be calculated the same way but with an additional charge. Given the measured m/z ratio for m_1 and m_2 the equations can be used to solve for the charge (Equation 1.3):

$$z = \frac{(m_2 - 1)}{m_1 - m_2} \tag{1.3}$$

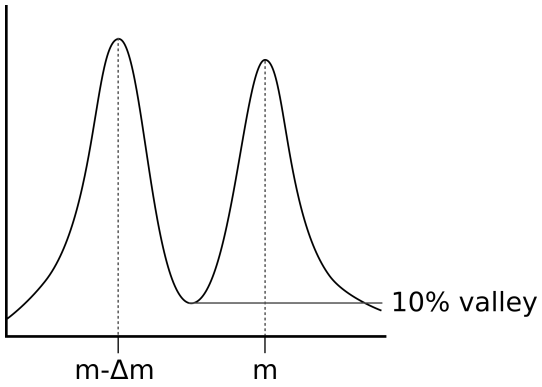
1.3 Mass Analysers

Once gas phase ions have been generated they are passed to a mass analyser which separates them based on their m/z ratio. Mass analysers are governed by different principles of separation however, they all make use of static or dynamic electromagnetic fields in isolation or in combination. Mass analysers can be divided into two main classes: scanning and non scanning analysers. Scanning analysers, such as the quadrupole, separate ions in time allowing only a particular m/z ratio to travel through the instrument at once. Non-scanning analysers, for example Time of Flight (ToF) analysers, allow simultaneous transmission of ions separating them in space.[24] As discussed by De Hoffmann and Stroobant [24] there are five characteristics upon which the performance of a mass analyser may be judged. These include: mass range, speed of analysis (scan speed), ion transmission, accuracy of measurement and resolution.

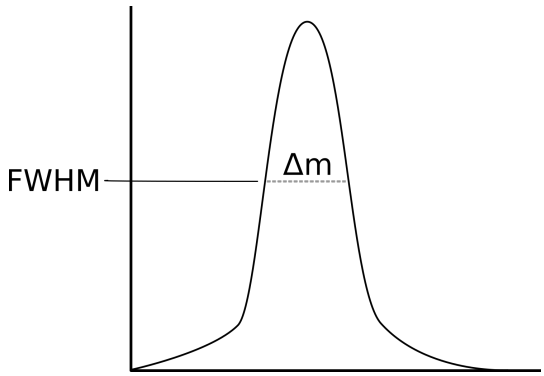
The mass range of an analyser is defined as the greatest extent of m/z measurement that can be determined. This is generally expressed in Thomsons (Th) or, for singly charged ions, in mass units (u). It is the upper limit of measurement quoted for an analyser. The mass accuracy of the analyser is the difference between measured and theoretical mass; most frequently expressed as Parts per Million (ppm), or in Daltons (Da). Scan speed is the speed at which the analyser can accomplish the measurement across a particular range. This is expressed in mass units per second (us^{-1}). Ion transmission is a dimensionless quantity and refers to the ratio of ions reaching the detector to ions which enter the mass analyser.

The resolving power of the mass analyser is a measure of its ability to separate signals from ions which posses a similar measured m/z and is thus descriptive of its ability to adequately resolve their mass. The precise definition of resolution has historically been the subject of debate.[98,9] Recently, it has been defined by the IUPAC Gold Book in two ways that allows the determination of resolution based on multiple and on single peaks: the 10% valley and the peak width definition respectively.[71] In both definitions resolution is defined as $\frac{m}{\Delta m}$, where m is the m/z recorded for the centroid of the peak. It is the difference in mass, Δm that is defined differently.

The 10% valley definition allows determination of resolution based upon the measurement



(a) Valley definition of resolution for multiple peaks. Δm is defined as difference in mass between the centroid mass of the two peaks m and m-Δm. 10% valley is indicated and represents 10% of the smallest peak height.



(b) Peak width definition of resolution for single peaks. Δm is defined as the peak width at half the peak height. FWHM is indicated and represents the full width at half the maximum height of the peak.

Figure 1.2: Definition of resolution for mass spectrometry as defined by.[71]

of two peaks, m and $m - \Delta m$ (Figure 1.2a). The peaks are considered resolved if the valley between them is less than or equal to a specified percentage of the intensity of the less abundant peak. For high resolution analysers such as FT-ICR mass spectrometers this is recommended to be 10%, for analysers such as the quadrupole and ToF a value of 50% is most frequently used. As Figure 1.2b shows the peak width definition or Full Width Half Maximum height (FWHM) allows determination of resolution for a single peak. Here Δm is defined as peak width at 50% peak height.

1.3.1 Quadrupole

Quadrupole analysers are scanning analysers which pass ions in successive intervals, separating them in time. This separation is accomplished by manipulating the trajectory of ions in oscillating electromagentic fields.[24] A quadrupole analyser consists of four parallel rods. A combination of RF (V) and DC (U) voltage is applied to the rods creating a hyperbolic EM field. As shown in Equation 1.4, 1.5 and Figure 1.3a the potential is applied to pairs of rods in an opposing manner such that each pair is negative of the other. Ions are pulled in a helical manner along the x and y direction by the influence of the electric field. If the ions have a stable trajectory they will move along the z-axis through the centre of the quadrupole. Ions with an unstable trajectory will discharge against the rods.

The potentials ϕ and $-\phi$ on pairs of rods are defined as:

$$\phi = U - V\cos\omega t, \tag{1.4}$$

$$-\phi = -(U - V\cos\omega t). \tag{1.5}$$

Where U and V are the DC and RF voltages respectively, ω is the angular frequency in radians per second and t is time.

The ratio of U and V can be manipulated to allow ions of a particular m/z to pass through the analyser on a stable trajectory (Figure 1.3b). In order to observe ions of consecutive mass in a successive fashion the constant DC voltage, U must be varied linearly as a function of the time-dependent RF voltage, V.[70] Quadrupoles are low resolution instruments.

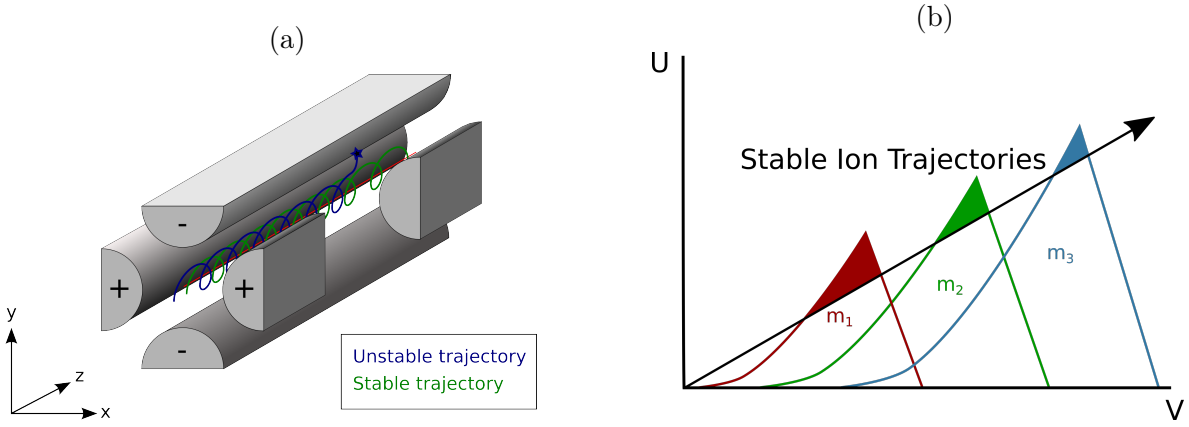


Figure 1.3: Principles of separation and stable ion trajectory through a Quadrupole mass analyser. a) Schematic showing ion trajectory through a quadrupole. b) Relationship between RF (V) and DC (U) voltage and stable trajectory. Arrow indicates the scan function or DC/RF. Filled triangles indicate stable trajectories for ions of three different masses where m3<m2<m1. Length of line in shaded areas indicates spectral peak area that will be generated.

As Figure 1.3b shows a quadrupole operated with a DC voltage set to $U = 0$ will have a resolution of 0. That is, all ions will have a stable trajectory. In this manner a quadrupole operating in RF only mode can transfer ions from one region of the mass spectrometer to another. The shaded areas in Figure 1.3b are directly related to the total ion count for a spectral peak of each m/z represented in the figure. This demonstrates the relationship between sensitivity and resolution. By increasing the slope of the scan function better resolution can

be achieved however, sensitivity will be decreased.

1.3.2 Time of Flight

The ToF analyser separates ions in space and transmits all of the ions to the detector at once. Packets of ions are separated in the ToF based on their path through the flight tube. The m/z ratio of the ions can be calculated by recording the time it takes for them to pass through the flight tube to the detector.[41] When the final ion, of the highest m/z, reaches the detector the flight cycle ends and another packet of ions can be pushed into the analyser.

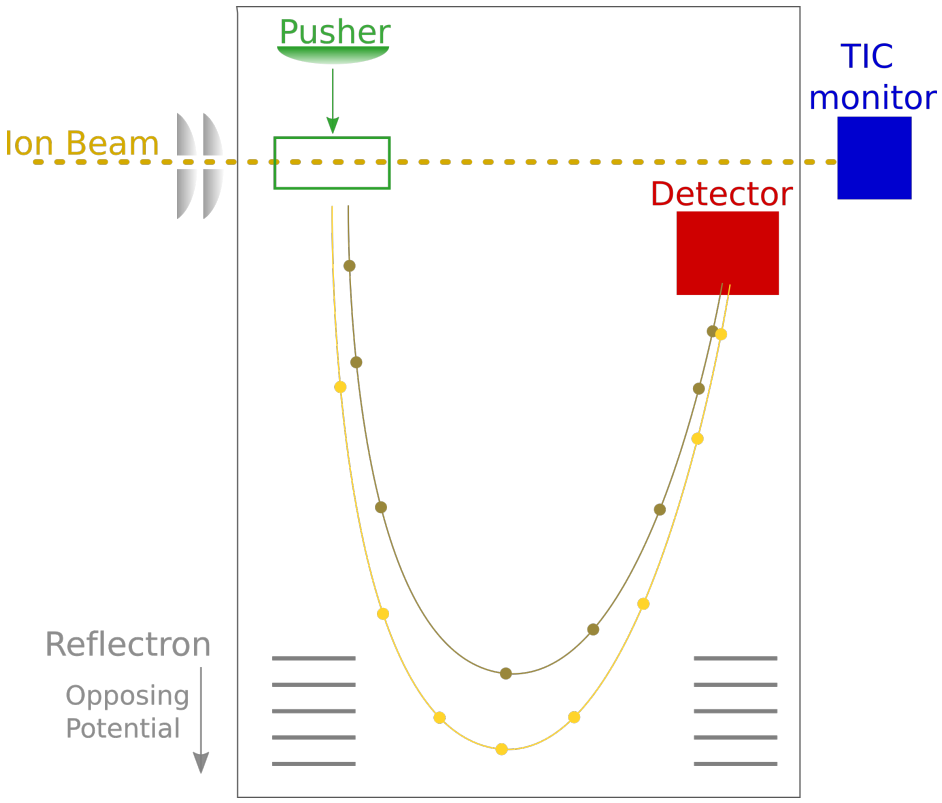


Figure 1.4: Path of a packet of ions through a ToF analyser. Ion beam is shown as a dashed tan line, high energy ion (yellow), low energy ion (brown). Pusher lens is shown in green and pulses a packet of ions from the ion beam into the analyser. In between pusher pulses the ion beam continues to the TIC monitor (in blue) where total ion count is recorded.

Historically, ToF analysers could only be used successfully with pulsed ion beam sources. Ion progression to the ToF was by means of axial injection through a narrow slit. With a continuous ion source blockages in this entrance caused a significant reduction in sensitivity.

In order to overcome this limitation Dawson and Guilhaus [23] developed an orthogonal acceleration ToF analyser (oa-ToF). The addition of a collimation lens focuses the ion beam whilst a pusher plate sited above the entrance to the ToF delivers a potential difference forcing the ions on an orthogonal trajectory to their initial direction of travel. While the pusher is accelerating ions into the ToF the ion beam continues on its original trajectory resulting in loss of signal. Recent advancements in pusher synchronisation have led to further increases in sensitivity. These are explored in Chapters 4 and 5.

Improvements in ToF technology also include the advent of the reflectron.[69] Ions entering the ToF have a range of initial kinetic energies. Hence ions with the same m/z will have a different flight path. Those with a higher energy have a longer flight path and consequently reach the detector at the same time as other ions with a lower m/z. The reflectron consists of an opposing electrical field that forces the ions back into the flight tube (Figure 1.4). Ions with a higher kinetic energy penetrate the field more deeply than those with lower energies. This increases the length of their flight path and refocuses ions with the same m/z equalising their arrival time at the detector plate.

ToF analysers have fast spectral acquisition rates (up to 20 spectra/s) which allows them to be compatible with Ultra-high Performance Liquid Chromatography (UPLC). This technique produces peak widths of 1-2 seconds.[113] The higher scan speed allows more of the ions to be utilised, leading to an increase in sensitivity. Continued advancements in ToF technology have enabled newer analysers in this class to reach resolutions of up to 40,000 FWHM.[6]

1.3.3 Orbitrap

Orbitrap mass analysers were introduced to the community in 2005 but was originally developed by Makarov [67]. The Orbitrap offers resolution comparable to FT-ICR instruments but at a substantial reduction in cost since they do not require a superconducting magnet. Orbitraps operate as both analyser and detector. Ions orbit around a central electrode inducing an image current that is interpreted by Fourier Transform (FT) analysis to provide both m/z and intensity values.

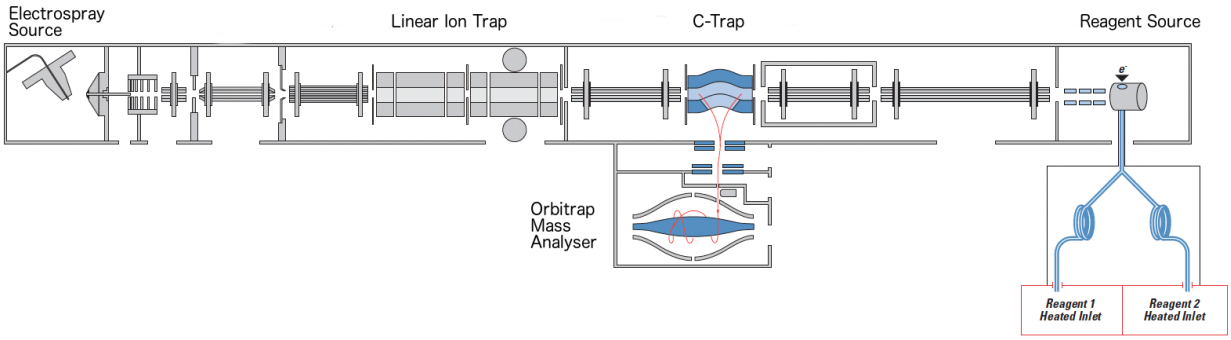


Figure 1.5: Schematic of Thermo Velos Orbitrap mass analyser. Reproduced with permission from Thermo Fisher Scientific [106]. Locations of ESI source, linear ion trap, C-trap, reagent source and the orbitrap mass analyser are marked. The ion beam through the C-Trap to the orbitrap analyser is shown in red. Ions rotate around the central spindle in the orbitrap analyser generating an image current.

The geometry of the outer and inner electrodes of the Orbitrap have been designed to create a quadro-logarithmic potential allowing ions to oscillate along the axis of travel in a sinusoidal fashion. As the frequency of these oscillations does not depend on the initial energy or spatial spread of the ions it may be used to determine their m/z. Due to the electrostatic nature of the field, packets of ions will continue to orbit the central spindle together in a helical manner. The total image current for the packet of ions is the sum of its constituent parts and thus the frequency spectra for each ion may be obtained using the fast Fourier transform.

A schematic of an example Orbitrap style analyser, the Thermo Velos, is shown in Figure 1.5. Following ESI ions are transferred to a linear ion trap (LIT) which serves to minimise space charging effects by regulating the number of ions entering the Orbitrap.[81] These space charging effects are common and are known to affect the image current. They are caused by the repulsive charges of ions shielding those closer to the centre. This is due to the proximity of the ions to the outer electrode.

Ions then progress to the C-Trap, a curved quadrupole type analyser which contains nitrogen gas at a pressure of approximately 0.1 Pa.[68] Due to this low pressure ions move back and forth along a path through the LIT and the C-Trap, confined by the gate and trap electrodes. Packets of ions are ejected orthogonally from the C-Trap and are focused into tight packets by a series of lenses to ensure a stable trajectory.

The use of such hybrid technology for the introduction of ions into the Orbitrap analyser enables resolutions of up to 60,000 FWHM and a scan speed of 1 scan per second. This type of analyser can scan over a mass range of approximately four orders of magnitude with an accuracy of 2-5 ppm. This advancement in methodology is discussed in more detail in Chapter 3. It should however, be noted that in the original cross-linking protocol paper it was only the precursor ions which were scanned in the Orbitrap analyser.[59] Fragmentation and subsequent analysis of the fragments ions was conducted in the LIT analyser at a much lower resolution.[87,12,73]

1.4 Ion Mobility

Ion Mobility (IM) is a gas phase separation technique based on similar principles to electrophoresis. Separation is achieved by exploiting differences in the transit time of species through an area of inert gas in the presence of an electric field. Mass spectrometry separates based on mass alone, however ion mobility isolates species based on their size, shape and charge. In doing so the method offers an extra degree of separation that can provide improved resolution of complex samples[75]

The original, and still widely utilised, method of IM separation employs a drift cell filled with a buffer gas using a low strength electric field. This field stimulates the migration of the ion through the drift tube. The technique requires that the energy an ion receives from the field is lower than that received through collision with the buffer gas molecules.[53] This method has the advantage that the collision cross section (CCS) of an ion can be directly measured from its drift time. As diffusion is the dominant force the ion velocity will be directly proportional to the field strength. This in turn is related to the mobility constant for the ion (Equation 1.6).[86]

$$K = \left(\frac{3q}{16N}\right)\left(\frac{2\pi}{kT}\right)^{\frac{1}{2}}\left(\frac{m+M}{mM}\right)^{\frac{1}{2}}\left(\frac{1}{\Omega}\right) \tag{1.6}$$

where q is charge state of the ion, N is the density of the buffer gas, k represents the Boltzman constant and T is temperature. m and M represent the masses of the buffer gas and the ion

respectively and Ω is its CCS value.

Traditional drift cells have the highest resolution, in the range of 50 -100, but low sensitivity. Ion transmission is poor due to the lengthy transit of an ion through the drift cell. Historically, most drift cell instruments were bespoke and not widely available.[20,10,45] The first commercially available mass spectrometer with the additional of ion mobility separation was produced by Waters Corporation in 2004 and incorporated Travelling Wave Ion Mobility.[38]

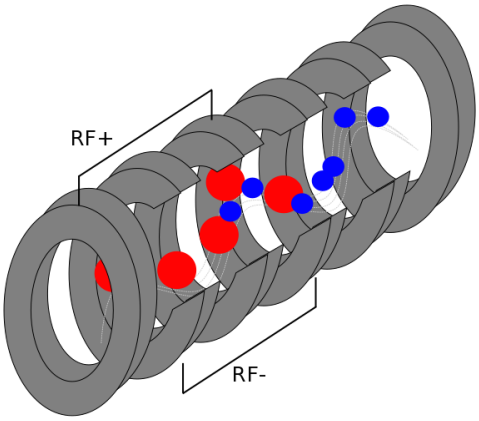
1.4.1 Travelling Wave Ion Mobility

Travelling Wave Ion Mobility separation (TWIM) makes use of a Stacked Ring Ion Guide (SRIG) filled with a buffer gas at low pressures. As shown in Figure 1.6 an opposing RF voltage is applied to consecutive ring electrodes creating a radially confining potential well.[38] A high strength field travelling DC voltage pulse is applied to ring electrodes in a directional manner so that the ions migrate in a wave like fashion through the centre of the IMS cell. Ions with lower mobilities fall behind the wave and exit the cell later than those that keep up with the velocity of the wave. Equation 1.7 defines the travelling wave velocity.

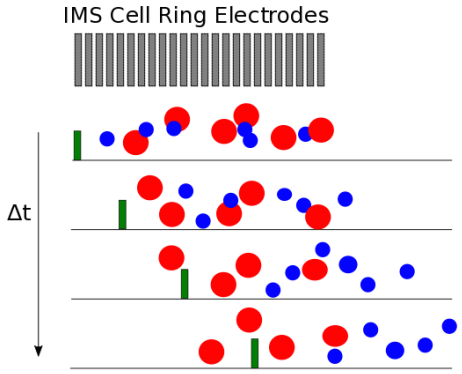
$$\frac{d_e}{t_p} \tag{1.7}$$

Where d_e is the distance between a pair of electrodes and t_p is the length of time the pulse is applied to the pair of electrodes. As wave velocity increases, the length of time the pulse is applied to a pair of electrodes decreases. TWIM is almost always coupled to ToF data acquisition. Once a packet of ions is released from the IMS cell a series of 200 pushes accelerates ions into the ToF recording 200 mass spectra in sequence. Upon release of each subsequent packet the process begins again.

A number of adaptations to early TWIM devices have been applied in order to improve both the sensitivity and resolution of the device. Nitrogen was found to give better separation of species than helium in a TWIM device.[36] As ions have a higher mobility in helium lower T-Wave amplitudes were needed to prevent the ions moving on a single wave. These lower amplitudes adversely affected the resolution of the device. The reduced mobility of ions in the presences of nitrogen allows greater wave amplitudes to be applied, giving better separation.



(a) Schematic of IMS cell showing separation of ionic species. Smaller ions shown in red exit the IMS cell first. Larger ions shown in blue. Alternating RF voltage contains the ions within the concentric rings.



(b) Representation of the travelling pulse in TWIM separation. DC voltage pulse shown as green block is applied to ring electrodes and forces the ions through the IMS cell like a wave. Smaller ions in blue larger ions in red.

Figure 1.6: Principles of Ion Mobility separation by Travelling Wave Ion Mobility.

However, higher pressures of nitrogen required a greater force to drive ions though the IMS cell. This caused scattering and fragmentation of the ions leading to a reduction in sensitivity. To improve sensitivity a helium curtain was added prior to the IMS cell, which helped to contain the nitrogen. An RF-only field was applied around the curtain allowing ions to gain higher mobilities as they enter the IMS cell. Finally, to improve resolution the length of the IMS cell was increased from 122 ring electrodes to 168 and the travelling pulse was applied to two pairs of electrodes rather than one. These changes were shown to improve resolution four-fold in the Synapt G2 instrument.[35]

In addition to the IMS cell, Synapt mass spectrometers also feature a further two SRIGs: the trap and transfer. These ion guides are located at either end of the IMS cell and operate between 7 - 9 μbar, equivalent to collision cell pressures. In addition to the confining RF voltage and the travelling wave DC pulse a third voltage can be applied to each ring in the trap and transfer SRIGs. An adjustable DC bias enables fragmentation of the ions facilitating tandem mass spectrometry.

1.5 Tandem Mass Spectrometry

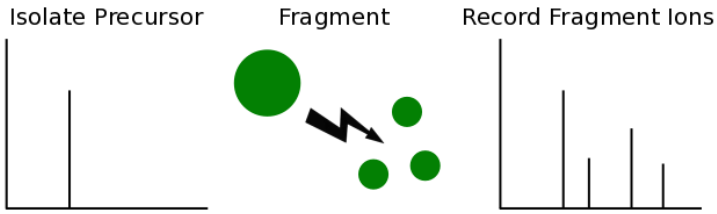


Figure 1.7: Representation of tandem mass spectrometry. Precursor ion masses are recorded and isolated generating MS spectra. Fragmentation occurs as represented by arrow. Fragment ions are generated from the isolated precursor and recorded by a mass analyser generating MS/MS spectra. This may be done in space or in time.

Tandem mass spectrometry (MS/MS) features two mass analysis steps separated by a dissociation step. Precursor ions are generated by the ionisation source and isolated by a mass analyser, where m/z and intensities are recorded. Following their initial analysis a fragmentation step occurs forming a second generation of ions. These are termed fragment ions. Both precursor and fragment ions are analysed in tandem to produce MS and MS/MS spectra. The basic principle of MS/MS is shown in Figure 1.7.

Tandem mass spectrometry can be performed in time or in space. Conducting the experiment in time requires the use of an ion storage device, such as a SRIG or an ion trap. Tandem mass spectrometry in space requires two physically distinct mass analysers with fragmentation occurring in between. Although multiple fragmentation methods exist the most commonly implemented method for the study of cross-linked peptides is Collision Induced Dissociation (CID).

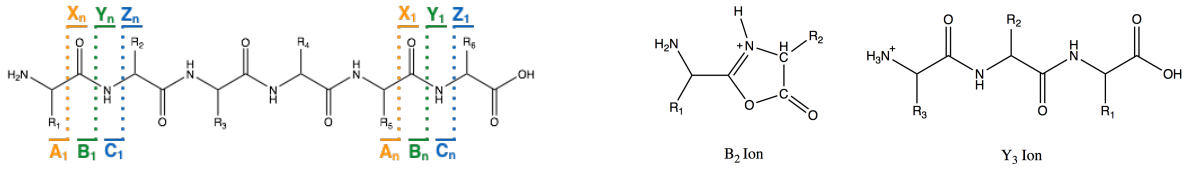
1.5.1 Collision Induced Dissociation

During collision induced dissociation ions are accelerated and collide with inert gas molecules.[49] Through a process known as collision activation, inelastic collisions occur between the analyte ions and gas molecules. A fraction of the kinetic energy is converted into vibrational energy allowing the most scissile bonds within the ion to break. The equation which describes the fraction of energy that can be converted is given in Equation 1.8.

$$E_c = E_{\text{lab}} \frac{M_t}{M_i + M_t} \tag{1.8}$$

Where E_c is the maximum amount of energy that can be converted, E_{lab} is the kinetic energy in the laboratory frame of reference. M_t and M_i are the mass of the target gas and of the ion respectively. Ions are accelerated into the collision chamber with higher energy collisions resulting from faster accelerations. The ratio of the mass of the target gas to that of the ion also affects the fraction of energy which may be converted. The energy is redistributed across the ion and fragmentation occurs at bonds lower in energy than the amount that is converted from the collision. The time it takes for the bonds to break is much slower than the speed of the collisions permitting ergodic ion dissociation.

The bonds in the peptide backbone are among the most scissile bonds, with the amide bond being the weakest. Consequently CID analysis of peptides yields repeatable patterns. Roepstorff and Fohlman [88] were the first to propose a uniform nomenclature for the observed fragmentation of peptides. This was later refined by Biemann [8] to give the nomenclature shown in Figure 1.8a.



(a) Roepstorff-Fohlman-Biemann fragment ion nomenclature.

(b) b and y ion formation after fragmentation at the R2 amide bond.

Figure 1.8: Principles and nomenclature of peptide fragmentation. A) Fragmentation at different bonds in the peptide backbone yields: A or X ions (orange), B or Y ions (green) and C or Z ions (blue). ABC ions are generated by fragmentation at the N' terminal side of the peptide. XYZ ions are generated through fragmentation at C' terminal side of the peptide. B) B ion formation proceeds through a cyclic oxazolone structure as shown. This prevents the observation of a B1 ion as two carbonyl groups are required.

The ions generated from fragmentation events along the peptide backbone are numerically labelled based on the site of cleavage within the ion. As such this labelling also refers to the number of residues contained in the fragment ion. Fragmentation at the peptide bond occurs through the localisation of charge to the lone pair in the nitrogen of the amide bond. The a, b, and c series ions are generated when the charge on the ion is localised at the N terminal of

the amino acid. For the x, y, z series, charge is localised at the C terminal. The mobile proton theory describes how this localisation of charge is permitted.[11] Ionising protons bound to basic amino acids may be transferred to each of the amide bonds. In this way a heterogeneous mixture of ions with different charge locations is generated allowing fragmentation at the sites indicated in Figure 1.8a. Although the b1 ion has been labelled in Figure 1.8a it should be noted that it is never observed. The b ion series consist of cyclic oxazolone structures which require at least two carbonyl groups to form (Figure 1.8b).

1.6 Cross-linking Mass Spectrometry

Cross-linking mass spectrometry has emerged as an important technique for the elucidation of protein structures that cannot be resolved by traditional methods. Cross-linking mass spectrometry (XLMS) provides a set of distance restraints that can be used in combination with electron microscopy (EM) and density fitting of subunit x-ray structures. These restraints guide the position and orientation of the individual components within EM density maps.[105] In this manner high resolution images of large and dynamic macro-molecular machines can be generated. Chemical cross-linking is also carried out in solution and is thus representative of the positions of amino acids in the native conformation of proteins or complexes.

Chemical cross-linking is straightforward in principle. A molecule with hetero or homo bifunctionality at either end is mixed with a protein of interest. The reactive group at either end of the cross-linker allows the molecule to form covalent bonds with specific amino acid side chains. Following cross-linking, the sample is digested and the resulting peptides are analysed by mass spectrometry. The highly accurate mass measurements generated during the MS/MS analysis are used to search sequence databases in order to elucidate cross-link positions. The hydrocarbon spacer arm between the two functional groups provides a distance restraint that can be used to triangulate the position of the two peptides in a wider three dimensional structure (Figure 1.9). This technique has been used successfully over the last twenty years to study complex systems such as the structure of the nuclear pore,[1] the protein components of the RNAP2 initiation complex,[73] the mitochondrial ribosome[42] and oxidative phosphorylation complexes within tissue.[18]

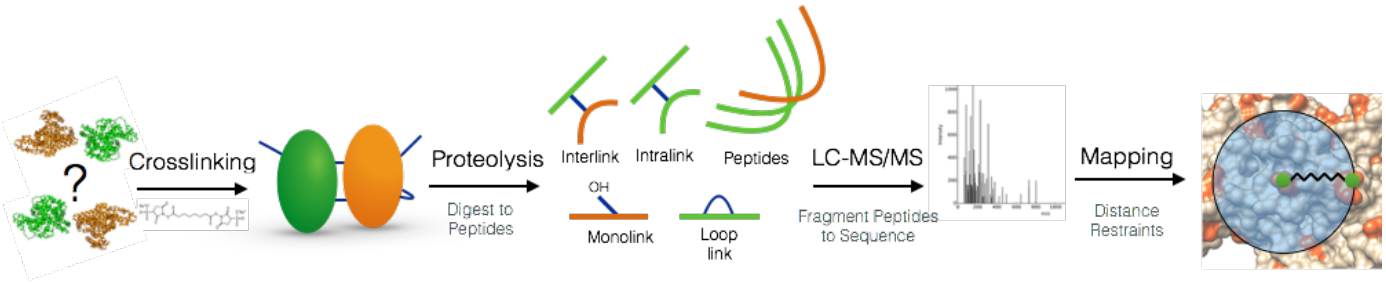


Figure 1.9: A cross-linking mass spectrometry workflow. Nomenclature as discussed in Leitner et al. [62]. Following exposure to the cross-linker the protein of interest is digested to produce a number of cross-linked products. Of these only the inter and intralinks are structurally informative. The cross-linked products are analysed by LC-MS/MS to sequence the peptides. Cross-links can then be mapped onto 3 dimensional structures or models to aid in structural determination and refinement.

XLMS must overcome some difficult obstacles before it can fulfil expectations.[64,46] These obstacles can be grouped into three broad categories: the experimental preparation of a protein or protein complex, the mass spectrometry techniques used to analyse the sample and the computational assessment of the raw data. From test tube to raw data set each stage of the analysis has inherent challenges. For single proteins and protein complexes the intralinks and interlinks respectively solely describe the structural topology. As a result of this imbalance much of the current literature of XLMS workflow development has emphasised the creation of novel cross-linking reagents.

1.6.1 Experimental Preparation of cross-linked Samples

Cross-linker Optimisation

Novel cross-linkers have been developed with the aim of increasing cross-link discovery prior to, during and subsequent to mass spectrometry analysis. The biotin tagged class offers the ability to isolate cross-linked peptides prior to MS analysis.[100] Leiker was developed to include a biotin tag and an azobenzene based cleavage site in addition to NHS Esters amino acid reactivity separated by a 9.3 Å spacer arm (Figure 1.10). This trifunctional cross-linker permits the addition of affinity purification to the cross-linking workflow. Cross-linked

peptides may be isolated from the digestion products through binding of the biotin tag to streptavidin beads. The cleavage site allows removal of the biotin tag by sodium dithionite. Removal of this tag is essential as it can interfere with LC-MS/MS analysis.[100]

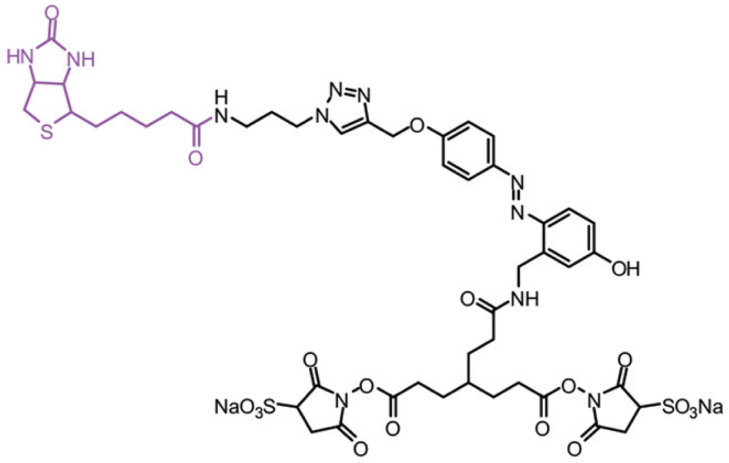


Figure 1.10: Biotinylated Azo-Leiker 1 (bAL1) cross-linker. Biotin group shown in magenta. Image produced using ChemDraw Professional version 16.0

The affinity purification adaptation has also been incorporated into the cleavable cross-linker class. The PIR cross-linkers developed by Tang et al. [102] enhance cross-link discovery during the MS analysis. PIR cross-linkers incorporate two CID cleavable bonds either side of a reporter ion. This system may also be attached to a biotin group allowing pre-analysis affinity purification. In this manner cross-link identification can be increased in two ways: by extracting modified peptides before analysis and by searching for a reporter ion signal during analysis.

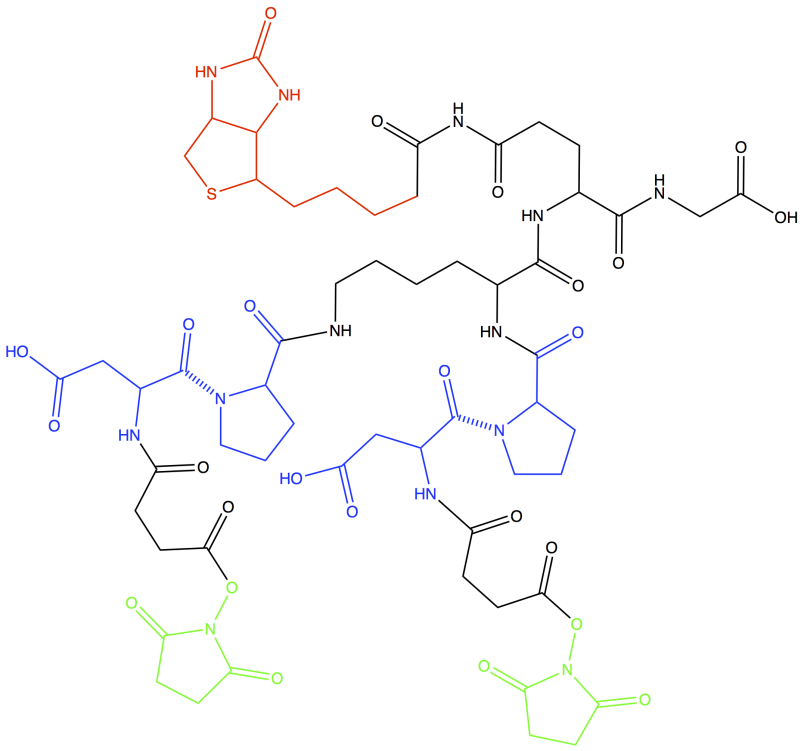
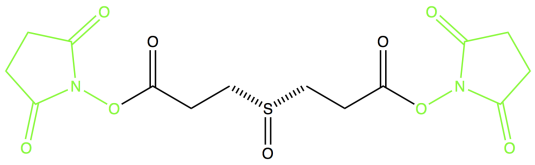
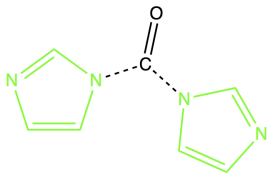


Figure 1.11: PIR cross-linker. Biotin tag shown in red, CID cleavable D-P shown in blue with dashed line representing scissile bond and leaving group shown in green. Image produced using ChemDraw Professional version 16.0

The most widely used cross-linker of this class features two D-P bonds either side of the reporter ion (Figure 1.11). These bonds have been shown to be susceptible to low energy CID cleavage[101,117] and fragment to reveal a diagnostic reporter ion, 771 m/z (MH+), that can be used to isolate spectra which contain cross-linked peptides. Following MS analysis of the precursor, CID at low energy reveals the reporter ion. Spectra containing this tag are recorded and can then be analysed using MS3. Alternatively an inclusion list of masses for precursor selection can be created and used to isolate cross-links in later LC-MS/MS experiments. By identifying the reporter ions cleavable cross-linkers can also aid in cross-link discovery during computational analysis. In addition, as this class of cross-linkers was synthesised using FMOC synthesis the linkers are composed of peptide bonds. Consequently they are treated as peptides by cellular machinery and are able enter the cell. This facilitates *in vivo* cross-linking, capturing the surrounding proteome of the protein in its native environment.



(a) Disuccinimdyl sulfoxide (DSSO) cross-linker. Leaving group shown in green, scissile bond shown as dashed line.

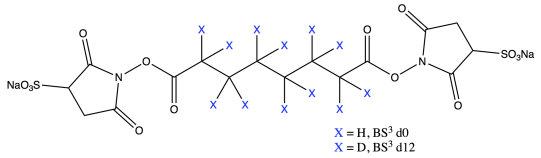


(b) 1,1'-carbonyldiimidazole (CDI) cross-linker. Leaving group shown in green, scissile bond shown as dashed line.

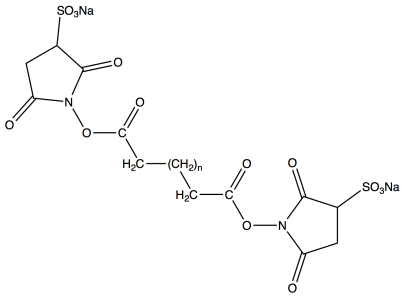
Figure 1.12: Schematic representation of cross-linker DSSO and CDI cleavable cross-linkers. Image produced using ChemDraw Professional version 16.0

The cleavable cross-linker disuccinimidyl sulfoxide (DSSO)[54] has a spacer arm of 10.1 Å and contains two labile C-S bonds which are again cleavable at low CID energies (Figure 1.12a). This allows each peptide to be isolated and sequenced separately by MS3, simplifying the spectra and enabling the use of standard proteomics software for peptides sequencing. Although these cleavable cross-linkers can aid in discovery they also require the use of MS3 and specialist scripts to recombine cross-link data.

More recently Hage et al. [43] introduced the first zero length cleavable cross-linker which fragments during MS/MS analysis. 1,1-carbonyldiimidazole (CDI) has a spacer arm of only 2.6 Å. This cross-linker can be used to gain information for the positions of amino acids in much closer proximity. The linker reacts with primary amines and has also been observed to connect these groups with the hydroxyl groups of tyrosine, serine and threonine. Upon collision activation CDI also undergoes fragmentation leaving behind a CO group on the primary amine of one of the generated fragments. This creates a doublet signal with a mass shift of 26 Da between the modified and unmodified fragment ion that can used to identify cross-links with analysis software. Unlike the DSSO cross-linker this fragmentation does not require MS3. It occurs at the same energy as peptide fragmentation, as a part of the MS/MS analysis.



(a) Molecular structure of BisSulfoSuccinimdylSuberate (BS3) cross-linker. Hydrocarbon spacer may be deuterated at positions marked with an X.



(b) Bis-sulfosuccinimidyl glutarate (BSG) cross-linker. Hydrocarbon spacer arm length can be varied and is shown in brackets.

Figure 1.13: Schematic representation of cross-linker BS3 and BSG cross-linkers that can be deuterated to create Heavy and Light Pairs. Image produced using ChemDraw Professional version 16.0

The most popular way to boost cross-link discovery without the need for specialist equipment uses isotopically labelled and unlabelled pairs of cross-linker.[74] By replacing the hydrogen atoms on the hydrocarbon spacer arm with deuterium a heavy version of the undeuterated cross-linker is generated (Figure 1.13). Both the light and heavy versions the cross-linker are used in equal concentration during an experiment. This generates pairs of heavy and light cross-linked precursors with a specific mass shift based on the number of deuterium atoms. Analysis software can then be used to isolate scan pairs possessing that mass shift as candidate cross-linked precursors. This class of cross-linker has the advantage that the mass shift can be controlled by the length of the carbon spacer arm. In addition a vast range of lengths and reactive chemistries are available. It can also be purchased in pre-mixed pairs of light and heavy cross-linker. The ease of use and diversity makes this technique one of the most flexible.

Almost all the cross-linkers mentioned above feature N-hydroxysuccinimide (NHS) esters that conjugate to the ϵ amine of lysine side chains. An example of this reaction can be seen in Figure 1.14. Although lysines have a high prevalence in most proteins and offer a higher reaction specificity than most amino acids, it has been shown that amino acids with hydroxyl groups in their side chains can also react with NHS esters. Kalkhof and Sinz [51] and Kalkhof and Sinz [52] found that cross-linking occurred not just between lysines but also between threonine, tyrosine and serine residues.

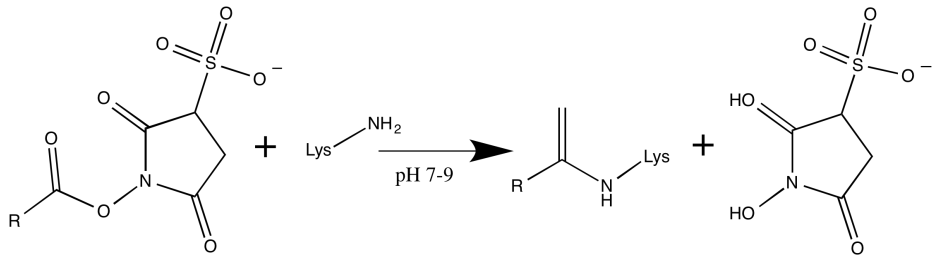


Figure 1.14: Reaction scheme for conjugation of NHS ester with a primary amine. Optimal pH for reaction is shown. Image produced using ChemDraw Professional version 16.0

Rappsilber [84] also raises concerns over the unrepresentative nature of data which relies upon specific residue reactivity as lysine residues are absent from areas involved in hydrophobic interaction. Furthermore, in conducting an analysis of 75 structurally resolved protein complexes Conte, Chothia, and Janin [22] concluded that lysine residues are often depleted at protein interfaces. This work is in agreement with the analysis conducted by Jones and Thornton [50]. For protein complexes, it may therefore be better to select a cross-linker that is not restricted by amino acid reactivity.

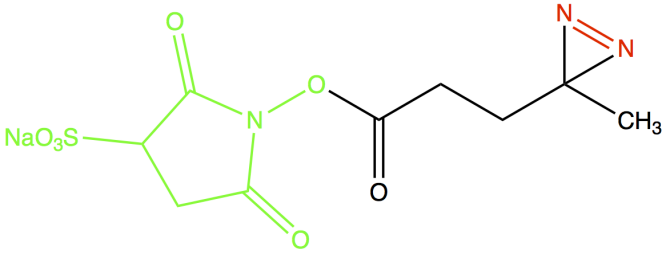


Figure 1.15: Sulfosuccinimdyl 4,4'-azipentanoate (sulfo-SDA) cross-linker. Leaving group following cross-linking shown in green. Leaving group following UV exposure shown in red. Image produced using ChemDraw Professional version 16.0

Belsom et al. [7] demonstrated the use of a photo-activated and highly promiscuous cross-linker to determine the structure of Human Serum Albumin (HSA). Sulfosuccinimidyl 4,4-azipentanoate (sulfo-SDA) is a heterobifunctional cross-linker that maintains the sulfo-NHS ester reactivity at one end but incorporates a diazirine group at the other (Figure 1.15). Upon exposure to UV light a carbene intermediate is formed that reacts non-specifically to any amino acid side chain or peptide backbone within 20 Å. Although the modelled HSA structure compared favourably to the crystal structure with RMSD of 2.5, 4.9 and 2.9 Å for

domains A, B and C respectively, the reaction of this diazirine group is so complete that it is difficult to interpret data and predict the exact position of the cross-link within the wider structure. Such refined detail is paramount for structural modelling approaches. As such, this non-targeted approach has limited applications.

Advancements in cross-linker chemistry are not limited to the selection above. Other classes exist that target both sulfhydryl and carboxylic acid moieties. All cross-linkers work with either homo or hetero-bifunctionality, allowing the user to optimise a workflow for a specific protein or complex.[94] It is therefore important to understand the sequence composition as well as the techniques that may be employed within the mass spectrometer in order to correctly tailor the choice of chemical cross-linker for an experiment.

Physiochemical Consequences of Cross-linking Peptides

The addition of most cross-link moieties modifies the structural environment hindering enzymatic digestion and thus encouraging sites of missed cleavage.[48] This conjugation also alters the distribution of energy along each peptide backbone making cross-linked peptides more difficult to fragment. As a result cross-linked peptides tend to be longer improving the likelihood of charge accepting amino acids within their sequence. The presence of two N termini on each peptide further permits the acceptance of charge, consequently cross-links carry a higher charge than linear peptides, most commonly $\geq +3$[33]

In order to generate sequence information a cross-linked sample is first digested. As depicted in Figure 1.9 this step produces multiple digestion products. As the number of cross-links in a sample is often an order of magnitude lower than the number of uncross-linked counterparts, enrichment processes are necessary to concentrate the sample and eliminate unmodified peptides. These methods take advantage of the differing physiochemical properties of linear and cross-linked peptides. They include the addition of size exclusion chromatography (SEC) or strong ion exchange chromatography (SCX) to the sample preparation.[60]

Although the larger size and highly charged nature of cross-linked peptides can be used as an advantage for enrichment, these characteristics also hinder annotation of MS/MS spectra. The presence of two peptides produces spectra that are more complex, containing a greater number of peaks at different charge states.[93] Furthermore, modifications due to digestion

such as oxidation of methionine residues and carbamidomethylation of cysteine residues also increase spectral complexity.[51]

1.6.2 Mass Spectrometry Analysis of Cross-linked Samples

A number of optimisations are possible for the analysis of cross-linked samples by mass spectrometry. These include, but are not limited to: varying collision energies and ramps, the type of analyser the fragment ion spectra are recorded in and how CID is conducted.

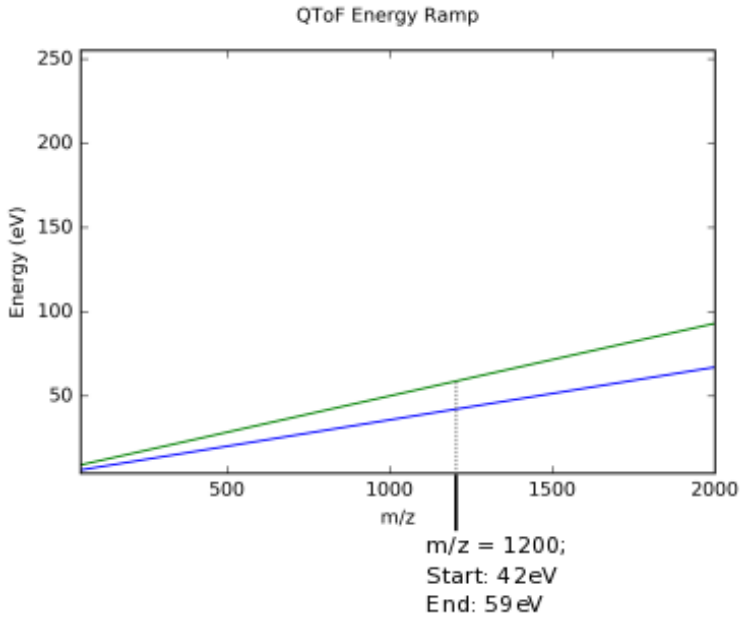


Figure 1.16: Example of collision energy ramping in a Synapt G2-Si. Low mass ramp shown in blue, high mass ramp shown in green. An ion of a particular m/z is exposed to the range of energies between the two ramps over the course of a scan. 1200 m/z is indicated on the image. Under these conditions an ion of this m/z will experience energies from 42 eV to 59 eV.

Collision energies in cross-linking experiments are often not reported. In Orbitrap analysers they are frequently set as the default 35% Normalised Colllision Energy (NCE). This energy setting increases linearly with the m/z recorded for the precursor under fragmentation and is equivalent to the HCD energy (eV) for an ion with a mass of 500 and charge of 1.[107] In a QToF collision energy is not reported as a dimensionless quantity, rather it is measured in eV. The fragmentation energy can be controlled by means of a ramp, where a range of

collision energies are delivered based on the precursor ion m/z. Figure 1.16 demonstrates this principle. Under the conditions displayed by this energy ramp an ion of 1200 m/z will experience a fragmentation energy that increases from 42 to 59 eV during the course of the scan.

Presently the most frequently reported method of acquisition method for the analysis of cross-linked peptides is Data Dependent Acquisition (DDA). This experimental design begins with a survey scan of the precursor ions entering the instrument. The m/z and intensities of the precursor ions are recorded. The most intense precursors from a scan are then isolated by the quadrupole and fragmented to provided an MS/MS spectra for the ion species. The maximum number of isolations depends primarily on instrument speed. At present the maximum achievable is approximately thirty separate precursor isolations per scan. The scan speed and maximum number of isolations is chosen *ab initio*. The DDA cycle ends once the maximum number have been analysed or when all detectable precursor ions have been sequenced. The process then repeats with another survey scan.

Due to the nature of the selection process DDA has an inherent bias for the most abundant precursors in a scan. The highly charged nature of cross-link peptides however, permits an adaptation to the DDA method that restricts precursor selection to ions with a charge state greater than +3. This prevents some singly or doubly charged ions from being selected, increasing sensitivity for higher charge state precursors. It should be noted that this optimisation reduces the signal from uncross-linked peptides but does not fully eliminate it. Tryptically digested linear peptides are able to carry a charge of +3. Hence DDA does not select solely cross-links present in a given scan.

1.6.3 Computational Analysis of Cross-linked Data Sets

In addition to experimental preparation and sample analysis the third aspect of a cross-linking workflow optimisation is the selection of analysis software. Interpretation of raw data generated by a cross-linking experiment is considered to be the greatest bottleneck in the workflow.[97,63] The software landscape is constantly evolving with new solutions appearing, at times, as frequently as the deprecation of older algorithms.

Several methods for the analysis of cross-linked datasets have been developed. They

fall into two broad approaches: a combination approach and a standalone method. In the former approach the user creates a database of all possible peptide combinations. These are linearised and a mass modification equal to the mass of the cross-linker is applied to specific residues determined by the cross-linker specificity. This approach allows the database to be searched by existing mass spectrometry analysis software. Traditional proteomics analysis software however, is not designed to look for highly charged peptides and deconvolution of the data to single charge is often a prerequisite.[77] The scoring algorithms utilised by such software are unsuitable for the analysis of cross-linked datasets since ions containing the cross-linker are not considered. This leads to lower scoring identifications with poorer confidence.[62]

In the latter and most popular approach dedicated cross-linking algorithms attempt to match the masses of observed precursors to the theoretical masses generated according to user inputs at the onset of the search. This may be achieved by considering the both peptides in the cross-link[87,40] or searching for one peptide and considering the second and the cross-linker as a modification.[116,111] In this case cross-links are identified by recombining peptide identifications that originate from the same MS spectra. Following a candidate match theoretical fragment ions are generated and matching algorithms try to annotate the peaks in the observed MS/MS spectra. Scoring algorithms subsequently attempt to measure the confidence of the identification.

The greatest challenge faced by analysis software is the complicated nature of cross-linked data sets. The presence of two peptides in a cross-link creates a search space of $\mathcal{O}(N^2)$ complexity. That is, a search space which increases quadratically for every peptide.[97,62,57] Consequently there are limitations on the number of proteins that can be analysed during the search.

Additionally, software developers must overcome intrinsic issues pertaining to the feature space. Cross-linking data has no "ground truth", due to the heterogeneous nature of cross-linked products, it is not possible to provide software with a set of known true positives. At present cross-links can only be identified through analysis of the data by existing software. Therefore cross-link identification is extrapolated on the basis of the scoring algorithms applied within the software. Any cross-link dataset generated in this manner and subsequently used as a training set will cause partitions in the data that are based upon the method of

identification in the original software.

To circumvent the issues of computational complexity and ground truth, a number of novel approaches have been developed. These include the creation of a set of known false positives,[87] the use of reporter ions either from a cleavable cross-linker[39] or from known fragmentation pathways[111,63] or by generating a set of known synthetic cross-linked peptides.[116]

The results of many of these attempts are complex algorithms which often lack interpretability. In addition implementation by end users often requires a high standard of requisite computational knowledge as instructions are either obfuscated or incomplete. Most solutions are platform dependent and many require the use of specific cross-linker chemistries.

1.6.4 xQuest

xQuest developed by Rinner et al. [87] and later improved by Walzthoeni et al. [114] is amongst the most longstanding computational solutions for XLMS analysis and has been widely use by the community.[78,58,109,73,17,79] xQuest/xProphet reduces the search complexity by making use of the mass shift observed in precursor spectra when light and heavy forms of an isotopically labelled cross-linker are used in combination.

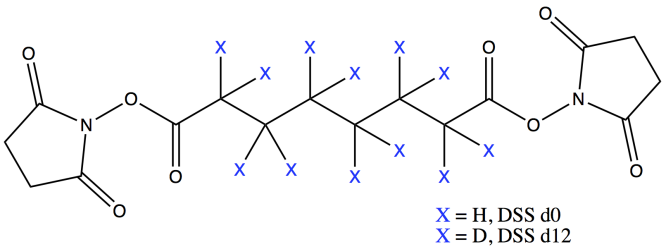


Figure 1.17: Molecular structure of DSS cross-linker. Cross-linker may be isotopically labelled. X represents Hydrogen (d0) or Deuterium (d12). Image produced using ChemDraw Professional version 16.0

xQuest searches a database of possible cross-links derived from protein sequence information. Precursor masses from the MS data which match the mass of candidate cross-links are termed peptide spectrum matches. xQuest requires both a heavy and light precursor to be present for a match to be confirmed. The quality of a match is determined using a series of scoring algorithms that provide a measure of confidence in the assignment. These scores are based

on correlation between theoretical and observed fragment spectra, probabilistic trial analysis and total ion contributions.

xQuest Scoring: intSum, WTIC, XCorr and matchodds

1.1 The most simplistic of the xQuest scores is simply the sum of the peak intensities for all peaks in the spectra. The score is described by Equation 1.9 where P in the peak intensity. The score is a measure of spectral quality.

$$\text{IntSum} := \sum_{i=1}^{n} P_i \tag{1.9}$$

The WTIC scoring function weights the observed total ion current (TIC) according to the length of each peptide in the cross-link. This score attempts to handle the problem of hybrid false positive identifications: in cases where a high score has been applied to a cross-link composed of a correctly matched short peptide and a longer decoy. It has the effect of reducing the TIC for larger peptides and increasing it for those that are shorter. The weight applied to the TIC for both the α and β peptide is defined by the original xQuest work as:

$$\text{wFrac}_{(\alpha/\beta)} = \frac{\text{NAA}_{\text{total}}}{\text{NAA}_{(\alpha/\beta)}} \frac{min_d}{min_d + max_d} \tag{1.10}$$

Where NAA is the number of amino acids, min_d is the minimum digest length and max_d is the maximum digest length. The final score is then computed by Equation 1.11.

$$\text{WTIC} = \text{wFrac}_{\alpha}\text{TIC}_{\alpha} + \text{wFrac}_{\beta}\text{TIC}_{\beta} \tag{1.11}$$

This score effectively re-weights the TIC but it is not specific to hybrid false positives. It is applied to all cross-links that have been identified. In essence, WTIC reduces the overall contribution of the TIC for larger peptides without false positive discrimination.

xQuest also features a correlation score adapted from Sequest.[66] The theoretical "best" spectra is first created for each peptide in the cross-link. This is a vector containing the m/z values of all possible b and y ion peaks (Figure 1.18). A second vector is created from the observed spectra and must be of equal length with unseen m/z peaks represented by 0.

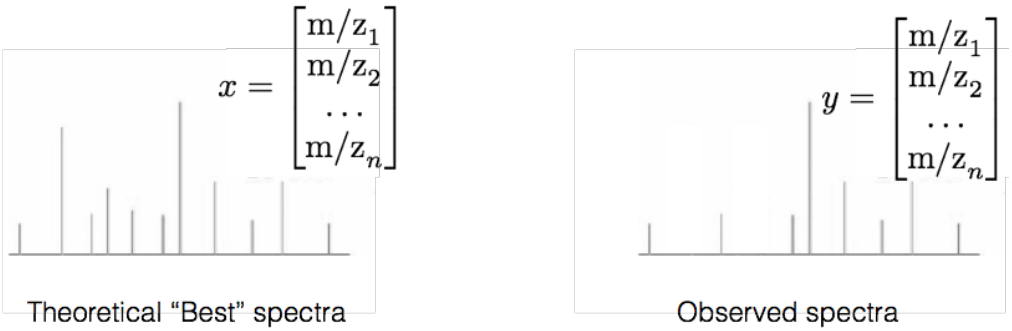


Figure 1.18: Calculation of inner product vector for XCorr score.

$$XCorr := \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \tag{1.12}$$

The current implementation of the XCorr score differs from that published in the original paper.[87] The scoring algorithm found within the software utilises normalised cross correlation (Equation 1.12), where x_i and y_i are $(m/z)_i$ values of peaks from the set of the theoretical and observed spectra respectively and $\bar{x}$, $\bar{y}$ and σ_x, σ_y are the means and standard deviations of the theoretical and observed values. This score is the normalised dot product of the vectors created from the observed and theoretical spectra. Negative values are permitted in the numerator when $x_i - \bar{x} < 0$ or $y_i - \bar{y} < 0$, but not both. An overall negative score is interpreted as poor correlation between the observed and theoretical spectra. XCorr scores are calculated for linear peaks found in both the heavy and light spectra and cross-linked peaks separately. Linear ions are generated from each individual peptide without the presence of the cross-linker whereas cross-linked ions contain the cross-linker.

A score based on a probabilistic Bernoulli trial completes the xQuest scoring algorithms. The MatchOdds score uses the cumulative binomial distribution function (Figure 1.19) to calculate the probability that a cross-link is genuine given: the number of theoretical fragment ions (n), the observed number of matches (k) and the mass accuracy (θ). It describes the probability of observing a number of ion matches between the theoretical and the experimental spectra, given the number of theoretical fragment ions. For a binomial distribution the probability of k successes in n trials is given by equation 1.13, where θ is the fairness.

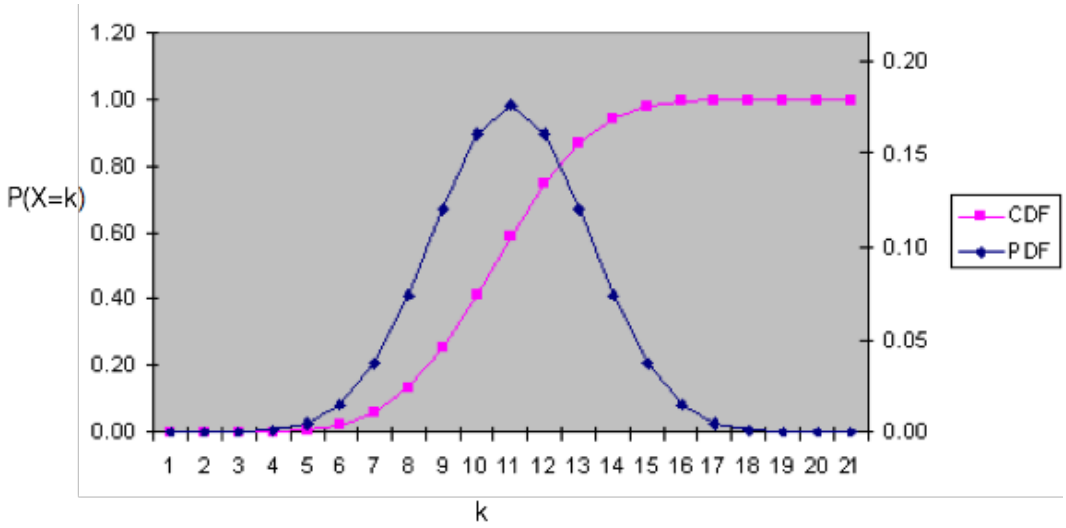


Figure 1.19: Representation of the Binomial probability density function (PDF) and the cumulative density function (CDF). K is the number of trials and the CDF is the sum under the curve for any point in the distribution.

$$\binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{1.13}$$

The binomial distribution is a good approximation for determining whether or not a cross-link is genuine as there can only be two outcomes: the cross-link is a true positive or the cross-link is a false positive. However, this distribution assumes that each trial is fully independent. The outcome of a trial does not depend on the outcome of any previous trials. In the case of identifying a cross-link, a trial is an ion match between the observed and theoretical spectra. This cannot be considered independent as the probability of matching an ion to the theoretical spectra changes depending on the number of previous matches.

xQuest Training: Linear Discriminant Analysis

In order to combine the five subscores into a single overall score that can discriminate between true and false positive cross-link identifications xQuest employs a statistical technique known as Linear Discriminant Analysis (LDA). The is part of of a broader class of supervised machine learning classification methods. In order to train this supervised classifier a dataset of known false positives was generated. Eight proteins were cross-linked separately with DSS, they were then digested and combined. This combined sample contained only intralinks and was

analysed by mass spectrometry with an Orbitrap mass analyser for the precursor and a LIT for the fragment ions. The resulting data files were searched using the xQuest/xProphet pipleine.

To further increase the quantity of known false positives a sequence database containing 100 random *Escherichia coli* proteins was used in the search. Any cross-links identified as an interlink between the eight mix and any inter or intralinks found containing *E. coli* proteins were known false positives. The test set comprised 370 true positives and 3040 false positives. These were characterised by strict criteria. True positives were identified as non-unique intralinks within the eight protein mix that scored above thirty with a ppm error when compared to the mass of the precursor between -4 and $+7$. The false positives contained interprotein cross-links with no score threshold.[114]

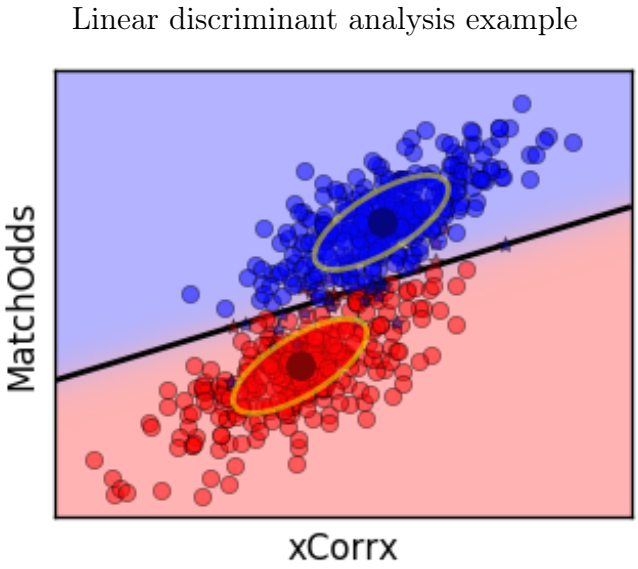


Figure 1.20: Example of separation by Linear Discriminant Analysis. Covariance of two subscores shown as yellow ovals, mean of each set as black dots. False positive in red, true positive in blue.

The five subscores MatchOdds, XCorrx, XCorrb, IntSum and WTIC were used as a co-ordinate system to plot the training set in five-dimensional space. The LDA was used to partition the true and false positive cross-link identification. Figure 1.20 illustrates this in a two-dimensional score setting. In the original work by[87] the LDA was cross-validated against the test set and provided a weight vector that could then be applied to each of the subscores in

order to differentiate between future cross-link identifications. Table 1.1 shows the calculated weights of the final contribution for each score and their equivalent mean contribution to final score.

Table 1.1: xQuest subscores: Weight derived from LDA and mean contribution to final score from training set (as calculated by Walzthoeni et al. [114]) MatchOdds subscore has the largest contribution to the overall final score.

Subscore	**LDA Derived Weight**	**Mean Contribution**
IntSum	0.018	27%
WTIC	12.82	3%
XCorrb	21.27	16%
XCorrx	2.48	2%
MatchOdds	1.9	52%

1.7 Aims and Objectives

Leitner, Walzthoeni, and Aebersold [59] published a protocol for the analysis of cross-linked proteins using an Orbitrap mass analyser. This high resolution analyser however, is only used to measure the precursor ions. Fragment ions are analysed in the LIT at a much poorer resolution. Presently higher resolution mass analysers are recommended for the analysis of cross-linked samples with HCD becoming the most dominant.[47] A tolerance of 5 ppm for matching precursor ions and 10 ppm for matching fragment ions is recommended for database searching. Such accuracy is not possible when using a LIT analyser, this can lead to an increased risk of missed or incorrect cross-link identifications.

Recent advancements in QToF technology have increased the achievable resolution providing fragment ion scans up to 40,000 FWHM. The scan speed of a ToF exceeds that of the Orbitrap which allows more ions to be utilised providing better sensitivity. Additionally, the four-fold resolution improvements to the IMS cell and the nature of the Triwave design in the Waters Synapt G2-Si provide an opportunity to incorporate ion mobility separation at different stages in the experimental design.

The separation of peptides into different charge states by ion mobility has been well documented.[83,96,103,45] Multiply and singly charged peptides are known to have different mobilities. As cross-linked peptides are both larger in size and more highly charged than un-modified peptides the addition of ion mobility separation should allow an increase in resolution resulting in a larger number of identifications with higher confidence.
The main objectives of my research were:

1. To implement a workflow for the analysis of cross-linking data generated on a QToF mass spectrometer from the test tube through to computational inspection using software developed by Rinner et al. [87], xQuest. To evaluate the effectiveness of this software when analysing QToF datasets.

2. To develop a cross-linking mass spectrometry workflow incorporating the use of ion mobility separation at both the precursor and fragment ion stages. Subsequently, to assess the effectiveness of each technique at improving the sensitivity and accuracy of cross-link determination for the structurally well-understood Bovine Serum Albumin (BSA).

3. To develop computational solutions that can incorporate the benefits of QToF data often not implemented by the currently available software in the field.

Chapter 2

Materials and Methods

2.1 xQuest Installation Requirements

A stand alone implementation of xQuest was installed on a Linux based workstation. The installation process requires intermediate web server deployment knowledge and command-line expertise. A full protocol for the installation has been compiled and can be found in Appendix A.

2.2 Preparation of Crosslinked Samples

BSA Crosslinking

A fully optimised protocol for cross-linking BSA has been previously developed within the Thalassinos lab.[82] Briefly, cross-linker concentration was titrated from 50 to 100 times molar excess and samples were extracted at various time points over the cross-linking experiment. A 100 times molar excess of cross-linker incubated over a thirty minute period was found to provide the highest number of cross-link identifications. This ratio of crosslinker to protein was used for all sample preparation unless explicitly stated within the text. Figure 2.1 shows an SDS PAGE analysis of a cross-linked BSA sample prepared as above alongside a control of uncross-linked BSA. The cross-linked sample appears as a group of higher molecular weight bands, confirming that cross-linking was successful.

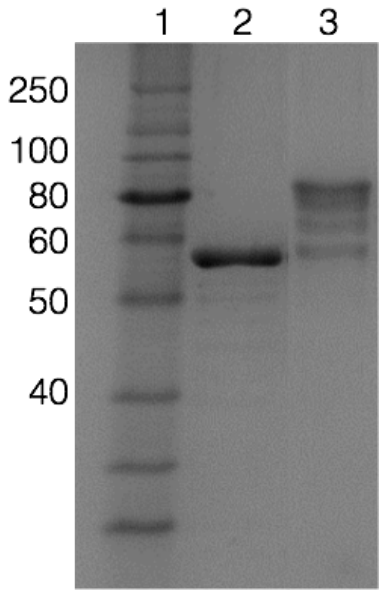


Figure 2.1: SDS PAGE results for 10 μM BSA samples. Lane 1) MW ladder, 2) BSA control with no cross-linker, 3) cross-linked BSA. Cross-linked BSA appears higher in mass with multiple bands representing different cross-linked oligomeric states.

0.3 mg/ml BSA (A7030, Sigma-Aldrich) and 1mg BS3 d0/d12 (Creative Molecule Inc.) were prepared in 20mM HEPES @ pH 7.6. 100 times molar excess of the cross-linker was added to the protein and the sample was then incubated at room temperature for thirty minutes under mild agitation. Following incubation the reaction was quenched by adding 1M Ammonium Bicarbonate to a final concentration of 50mM. The samples were then evaporated to dryness in a Thermo Savant speed vacuum.

9 Protein Mix Crosslinking

The following proteins were purchased from Sigma Aldrich and crosslinked as below:

2 mg/ml of each protein above were prepared in 20mM HEPES @ pH 8.2. 1 mg DSS d0/d12 was prepared in anhydrous DMF (Creative Molecules 001S). The crosslinker was added to the protein and diluted to a final concentration of 2.5mM DSS d0/d12. The sample was then incubated at room temperature for thirty minutes under mild agitation. Following incubation the reaction was quenched by adding 1M Ammonium Bicarbonate to a final concentration of 50mM. 25μg of each XL protein was pooled to form a single sample of 200 μg combined crosslinked protein. The sample was then evaporated to dryness in a Thermo

Table 2.1: List of monomer proteins in 9 Protein Mix with Uniprot ID

Uniprot ID	Protein Name	Species
P00432	Catalase	Bovine
P00563	Creatine Kinase	Rabbit
P00883	Aldolase	Rabbit
P24627	Lactotransferrin	Bovine
P02789	Ovotransferrin	Chicken
P02769	Serum Albumin	Bovine
P81461	Concanavalin A	Maunalua
P68082	Myoglobin	Equine
P00330	Alcohol Dehydrogenase	Saccharomyces cerevisiae

Savant speed vacuum.

In Solution Digestion and Solid Phase Extraction

The cross-linked protein mixture was resuspended in 8M urea at 1.1mg/ml concentration. 1% Rapigest (Waters, 186001860) to a final concentration of 0.1 % was added to aid solubilisation before digestion. The sample was then incubated with 10mM Dithiothreitol at 37°C for thirty minutes to denature the protein and reduce the disulphide bonds. Following incubation the sample was allowed to cool to room temperature. In order to prevent the reformation of disulphide bonds iodoacetamide was added to the denatured cross-linked protein sample. This was added to a final concentration of 20mM. As iodoacetamide is unstable when exposed to light the mixture was incubated in the dark, at room temperature for thirty minutes.

The sample was then diluted with 50 mM Ammonium Bicarbonate to reduce the final concentration of Urea to $<$ 1M. Trypsin (Promega, sequencing grade) was added to the sample to a ratio of 50:1 protein to enzyme. The reaction was incubated over night at 37°C with mild agitation. Following over night incubation enzymatic activity was quenched by adding formic acid to a final concentration of 2% (vol/vol). In preparation for SEC fractionation the sample was purified using SPE(50 mg Sep-Pak c18 (Waters, WAT054960)).

SPE cartridges were washed with 500 μL MS Grade acetonitrile(ACN) then equibrilated twice with a wash solution of 95% MS Grade H_2O 5% MS Grade ACN 0.1% MS Grade Formic Acid. The sample was loaded onto the cartridge and washed a further two times.

Cross-links and peptides were eluted twice with 50% MS Grade H_2O 50% MS Grade ACN 0.1% MS Grade Formic Acid. The purified sample was then dried down for Size Exclusion Chromatography (SEC) separation.

MicroAkta Size Exclusion Chromatography Fractionation

Following SPE the sample was resuspended in 20 μl of SEC buffer (degassed water/acetonitrile/TFA at 70/30/0.1 vol/vol/vol). 15 μl of sample was injected onto an equilibrated Superdex Peptide PC 3.2/3.0 column (GE Healthcare, Part no. 17-1458-01). 100 μl fractions were collected. Previous investigations revealed that with high reproducibility, fractions A12 to B3 contain the most cross-links for BSA (Figure 2.2). The digested, crosslinked, fractionated samples were then evaporated to dryness and resuspended in 250 μl Liquid Chromatography Mass Spectrometry (LCMS) buffer; 95% H_2O, 5% acetonitrile and 0.1% formic acid. All solvents must be MS grade.

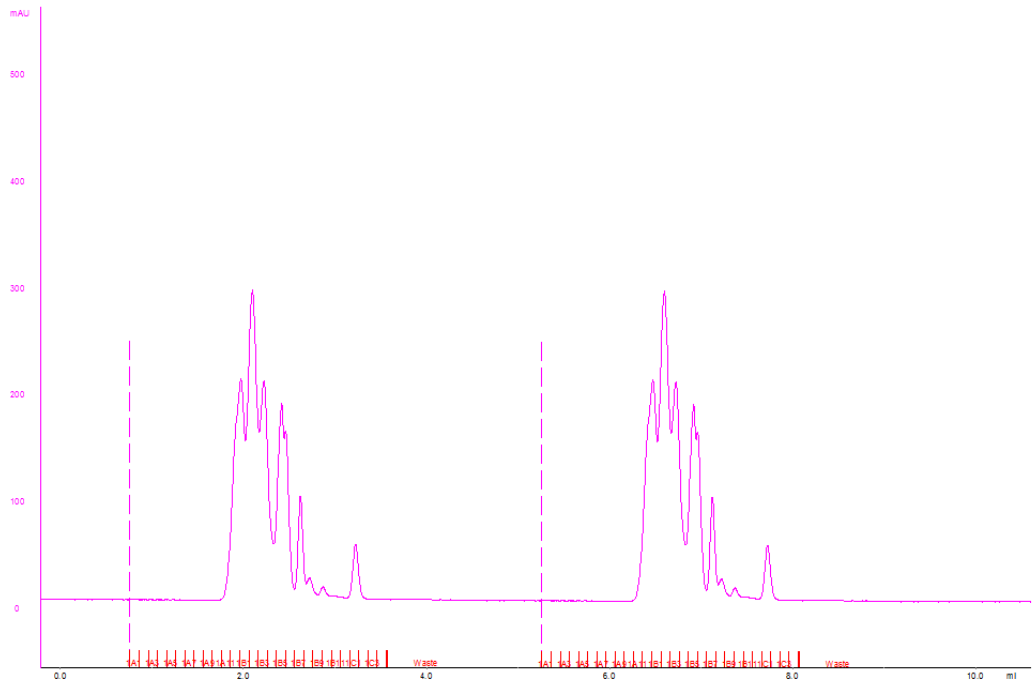


Figure 2.2: SEC trace for BSA digest. A high level of reproducibility is shown across repeated biological runs.

2.3 LC-MS/MS Analysis

Samples were introduced using nano-Ultra Performance Liquid Chromatography (10kPsi nanoacquity Waters Corp. Milford, MA, USA) with buffers: MS Grade water, 0.1% formic acid (A) and MS Grade acetonitrile, 0.1% formic acid (B). Samples were desalted by a reverse-phase SYMMETRY C18 trap column (180 μm internal diameter, 20 mm length, 5 μm particle size, Waters Corp.) at a flow rate of 8 μl/min for three minutes in 99% Buffer A. Peptides were then separated using a linear gradient (0.3 μl/min, 35°; 97-60% Buffer A over 90 mins) using a BEH130 C18 nano-column (75 μm internal diameter, 400 mm length, 1.7 μm particle size, Waters Corp.).

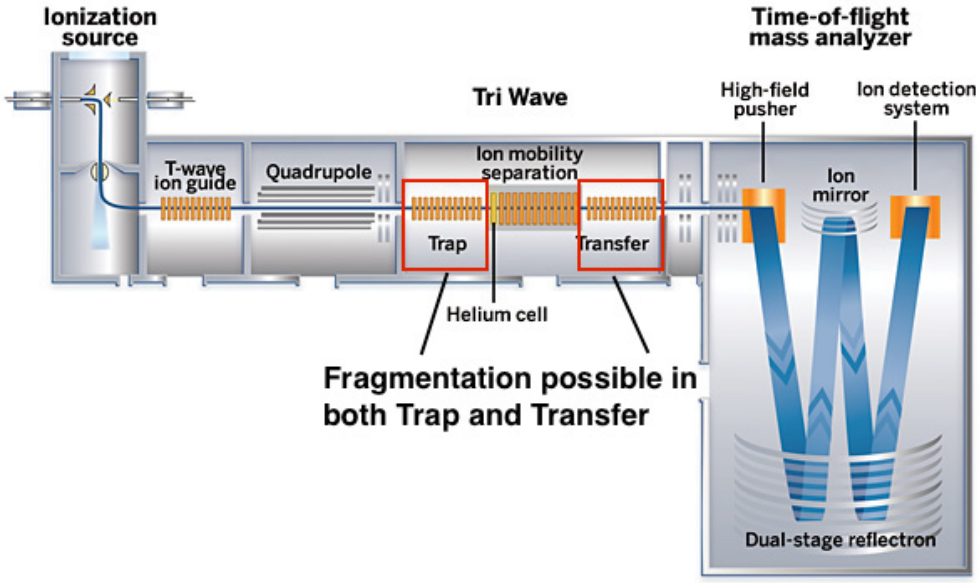


Figure 2.3: Schematic representation of Waters Synapt G2-Si Quadrupole Time of Flight mass spectrometer. Sites of possible peptide fragmentation are indicated.

The LC was coupled to the Water Synapt G2 Si quadrupole time-of-flight mass spectrometer (Figure 2.3). The ToF analyser was externally calibrated from m/z 175.11 to 1285.54 using [Glu^1]-fibrinopeptide B (Sigma aldrich) at 500 fmol/μl, hereafter referred to as GFP. Data were post acquisition lock-mass-corrected using the monoisotopic mass of the doubly charge GFP precursor at 785.84 m/z. Lock spray was delivered by an Auxillary Solvent Manager into the NanoLockSpray interface and sampled every 60 seconds. Accurate mass measurements were made using Data Dependent Acquisition over a mass range of 50-2000Da

with a scan time of 0.15 s and an interscan delay of 0.05s. Following collision energy testing (Section 3.3.4) collision energy was ramped according to m/z using the following parameters: LM 10-20 eV, HM 30-60 eV. Data was acquired in resolution mode for precursor scans with improved sensitivity for fragment ion scans. These parameters were maintained throughout the study unless explicitly stated within the text.

2.4 Raw Data Processing and Cross-link Analysis

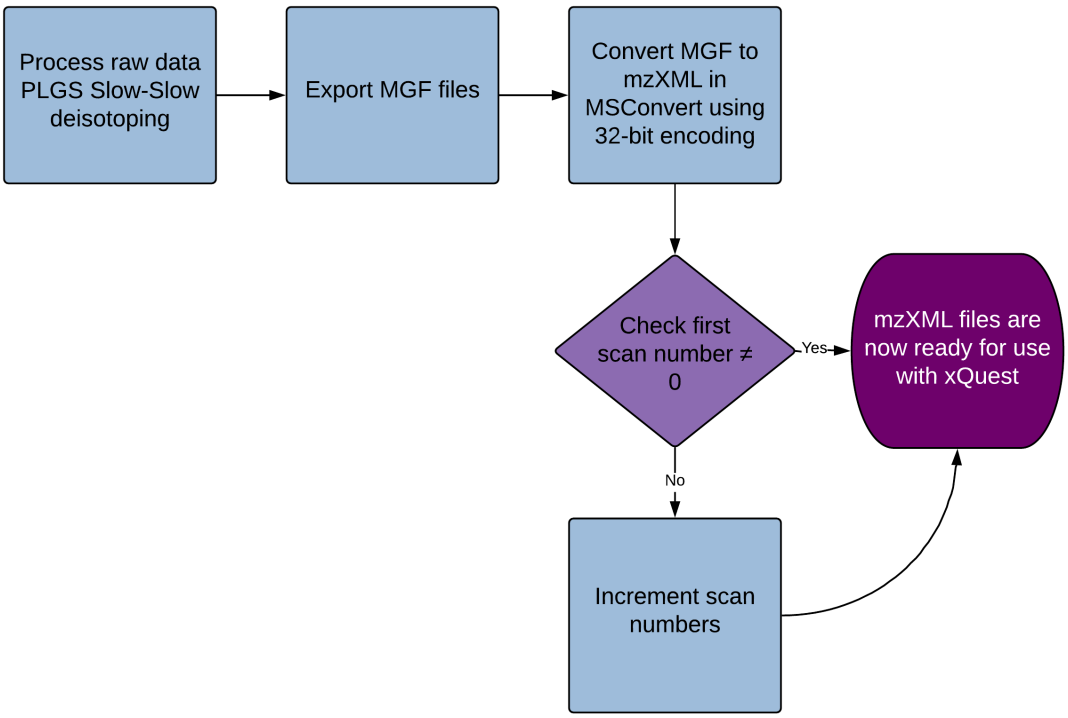


Figure 2.4: Data formatting pipeline for use of xQuest cross-linking analysis software with QToF data. Steps to process raw data and convert MGF files are shown.

Waters raw files were processed in Protein Lynx Global Server (PLGS) v3.2.0 (Waters Corp.) The methods for processing are explained in detail by Geromanos et al. [31]. Briefly this software makes use of the Savitzky-Golay smoothing algorithm[91] to increase signal and reduce noise in the data. PLGS also carries out peak detection, deisotoping and deconvolution to provide centroid data for the monoisotopic peak of peptides.

After PLGS processing, the files are exported in Mascot Generic Format (MGF) and then further converted to mzXML format with 32 bit encoding as required for xQuest input. Conversion to mzXML was accomplished by MSConvert from proteowizard 3.0.7414.[56] Figure 2.4 shows the data formatting pathway.

The final processing parameters were as follows for the precursor and fragment ions; lock mass calibration using the monoisotopic doubly charge peak of GFP with a tolerance of 0.25 Da, noise reduction of 35% and deconvolution using the slow algorithm with thirty iterations. The final mzXML files were then analysed for the presence of cross-links using a stand-alone implementation of xQuest installed on a workstation operating with Ubuntu 14.04. A full description of the installation process in given in Appendix A.

xQuest was designed and trained on sample data obtained from an Orbitrap mass analyser, as such modifications to the search parameters were necessary to accommodate QToF data. Essential modifications are displayed in Table 2.2. Briefly; data obtained from an Orbitrap mass analyser differs from data obtained in a QToF as it is not deconvoluted and contains multiple charge state fragment ions. In addition, the mass range of Linear Ion Trap analysers, as used in the original work,[87] are limited to between 200 and 2000 Da. QToF data has a maximum range of between 50 and 5000 Da.

Full search parameters can be found in Appendix B and include two possible missed cleavages, variable modification of methionine oxidation and a fixed modification to cysteine residues caused by carbamidomethylation as a result of the additional of iodoacetamide.

Table 2.2: Alterations to xQuest.def file for use with QToF mass spectrometer. Default values for xQuest parameters as stated in literature are shown along with modifications made to incorporate QToF style data.

Parameter	**Value type**	**QToF Value**	**Literature value**	**Description**
ioncharge_common	Integer	1	1, 2, 3	Charge states to be considered for common ions
ioncharge_xlink	Integer	1	2, 3, 4, 5	Charge states to be considered for crosslinker containing fragment ions
minionsize	Integer	50	200	Minimum ion size in MS2 mode to be considered
maxionsize	Integer	5000	2000	Maximum ion size in MS2 mode to be considered

2.5 Computational Analysis

Unless otherwise stated in the text all computational analysis was conducted using Python 3.5.0.[89] Table 2.3 contains details of all libraries and versions implemented in the code.

Table 2.3: List of Python Libraries and Versions

Library	**Version**
biopython	1.68
et-xmlfile	1.0.1
matplotlib	1.5.3
matplotlib-venn	0.11.4
numpy	1.11.2
pandas	0.22.0
seaborn	0.7.1

Chapter 3

Analysis of Cross-links identified by xQuest/xProphet in QToF Data

3.1 Introduction

Cross-linking mass spectrometry (XLMS) can be used to gain insights into the structure of proteins and complexes that are large in nature or posses high levels of dynamics and flexibility.[84,105] Cross-linking Mass Spectrometry (XLMS) provides a set of distance restraints that describe the relative position of two amino acids in a wider three dimensional structure.

The earliest cross-linking workflows incorporated LTQ-Orbitrap mass analysers. Most protocols specified that Collision Induced Dissociation (CID) and subsequent analysis of the fragment ions was performed using a Linear Ion Trap (LIT).[87,73,12] The lower resolving power of this analyser however, can lead to incorrect peptide annotations and an increased risk of false positive cross-link assignment. The accuracy of this analyser is now recognised as insufficient to adequately annotate fragment ions in complex cross-link spectra.[47] High energy Collisional Dissociation (HCD) in which fragment ions are re-injected into the Orbitrap is now the method of choice for cross-link analysis as it offers high resolution MS and MS/MS data.

Currently, almost all cross-linking studies utilise Orbitrap mass analysers. As a result, the majority of available cross-linking software has been optimised for Orbitrap style data, that is, centroided data with multiple charge state for fragment ions. As previously described in

Section 1.6.4, the xQuest/xProphet application[114] is no exception. Full details of the xQuest scoring algorithms are provided in the Introduction of this work. Briefly, xQuest uses linear discriminant analysis to weight the results of several subscores from spectra that contain candidate cross-links. In order to adapt the software for the analysis of data collected on a Waters Synapt G2-Si mass spectrometer differences in instrument operation must first be considered.

Recent advancements in Time of Flight (ToF) technology have led to an increase in achievable resolution, providing fragment ion scans up to 40000 FWHM.[6] In addition, ToF resolution does not depend on acquisition time, offering a near constant resolution across the mass range and faster scan speeds compatible with Ultra-high Performance Liquid Chromatography (UPLC) for both precursor and fragment ion scans.[80,118] Furthermore, the ability to seamlessly integrate QToFs with Ion Mobility Separation (IMS) offers a potential extra degree of separation by size, shape and charge, without the requisite need for additional analysis time.[83] In addition to providing an extra degree of separation the milli-second time scale of the technique sits effectively between LC separation spanning seconds and ToF analysis which spans the micro-second range.[44] The adaptation of such techniques for cross-linked peptides is discussed in more detail in later chapters. IMS may provide an opportunity to increase cross-link yield through exploitation of the larger and more highly charged nature of cross-linked peptides.

As well as the differences in functionality outlined above there are fundamental differences in how CID fragmentation is conducted by each type of instrument. Orbitrap mass analysers offer a range of fragmentation options including CID and beam type HCD. CID, conducted in the LIT analyser, is performed at 5 mBar with helium as the preferred target gas.[81] A QToF operated with argon rather than helium allows higher amount of potential energy to be converted to vibrational energy and thus increases fragmentation efficiency (Figure 3.1). In addition, the trap cell in the Synapt operates at $8.94x10^{-3}$ mBar.[35] Fragmentation during a HCD experiment is performed in a separate collision cell and are stored in the C-Trap so that the fragment ions may be passed back to the Orbitrap analyser. Operation of the HCD collision cell recommends the use of argon rather than helium as a collision gas. The typical pressure for the Thermo Fusion C-Trap is reported to be $13x10^{-3}$ mBar.[76] This

difference in operating pressure changes the number of collisions to which the ions are exposed. Furthermore, the efficient fragmentation of precursor ions in an Ion Trap can only occur at energies that limit the mass range of trappable ions. Ions able to form stable trajectories through the analyser fall within the range of 200-2000 m/z.[65] HCD and QToF CID do not suffer from this limitation and, with the right calibration, can scan between 50-5000 m/z.

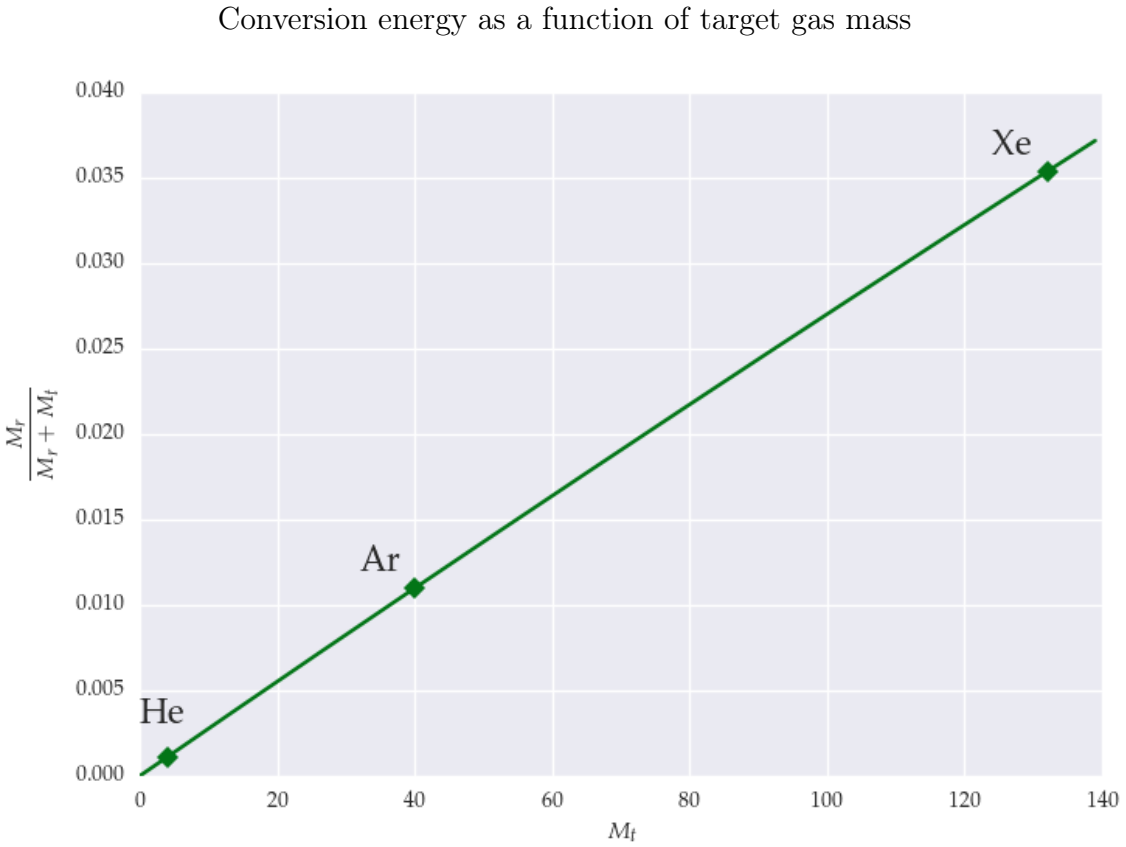


Figure 3.1: Energy available for conversion as a function of target gas mass. A mass of 3080 Da has been used to represent a cross-linked peptide based on the following assumptions; tryptic digests produce peptides with an average length of 14 amino acids,[14] with an average molecular weight of 110 Da per amino acid and including the presence of two peptides. Mass of the cross-linker has not been considered.

Although a robust protocol for the use of a QToF instrument has yet to be published, much work has been done to optimise the fragmentation techniques and collision energies using both hybrid[30] and tribrid[32,57] Orbitrap mass analysers. These studies concluded that the most optimal method of fragmentation for cross-linked peptides is HCD. Kolbowski, Mendes, and Rappsilber [57] further concluded that whilst fragmentation efficiency reaches

a maximum in the range of 22-24% normalised collision energy (NCE) when using CID techniques, HCD achieved maximal fragmentation efficiency between the range 26-30% NCE.

The differences in instrument parameters and the observed energy dependant nature of fragmentation efficiency motivates the optimisation of collision energy for cross-link analysis by a QToF mass spectrometer. Here we present a method for the analysis of cross-linked peptides analysed using a QToF geometry. We performed triplicate analysis of 6 different energy ramps in order to determine the optimal energy for cross-link fragmentation. The resulting approach allows QToF data to be analysed using existing cross-linking software, xQuest,[114] with minimal adaptations. We analyse the effects of collision energy on the fragmentation efficiency of Bovine Serum Albumin (BSA) cross-linked with isotopically labelled bis(sulfosuccinimdyl)suberate (BS3d0d12). Finally, in contrast to data collected from Orbitrap analysers[111,33,57] we also demonstrate improved fragmentation of the smaller peptide in the cross-link.

3.2 Materials and Methods

Full details of sample preparation are given in the Chapter 2: Materials and Methods. Mass spectrometry analysis was conducted in accordance with this protocol whilst encompassing the following alterations to the collision energy described in Figure 3.2.

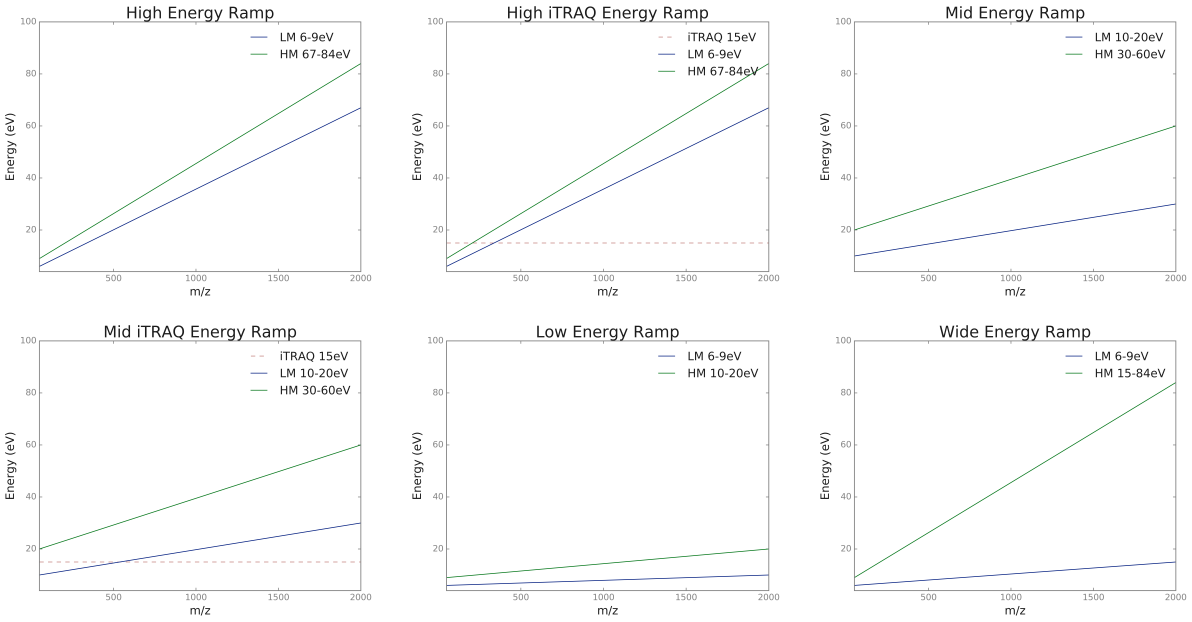


Figure 3.2: Energy ramps tested during parameter optimisation. An ion of a particular m/z is exposed to the range of energies between the LM (blue) and HM (green) ramps. For more information see Figure 1.16.

Ramps were tested in the same daily period to minimise abnormalities caused by fluctuations in solvent composition and temperature. Unless otherwise stated the order was "High", "HighiTRAQ", "Mid", "MidiTRAQ", "Low", "Wide". Blank runs were included to prevent cross-link carry-over between experiments. Ramps were tested in triplicate using the same sample of cross-linked BSA.

The "Mid" ramp is the manufacturer-recommend ramp for fragmentation of linear peptides in the Synapt G2-Si. The "High" ramp was originally designed for use in a method know as HD-DDA.[44] This method is explained in more detail in Chapter 5. Briefly, HD-DDA synchronises the pusher in the ToF to the mobility pattern of peptide fragment ions in order to reduce the duty cycle of the instrument. As future method development work incorporating ion mobility separation was desirable this ramp was also tested.

Previous studies of lysine-lysine DSS intramolecular cross-links revealed fragmentation of the amide bond within the cross-linker itself.[48,111] In order to assess the effects of energies on fragment ions containing the cross-linker three further ramps were considered; a "Low" energy ramp and an iTRAQ modification, which was added to the Mid and High ramps. The iTRAQ method was originally designed for quantitation of differentially labelled proteins. The isotope

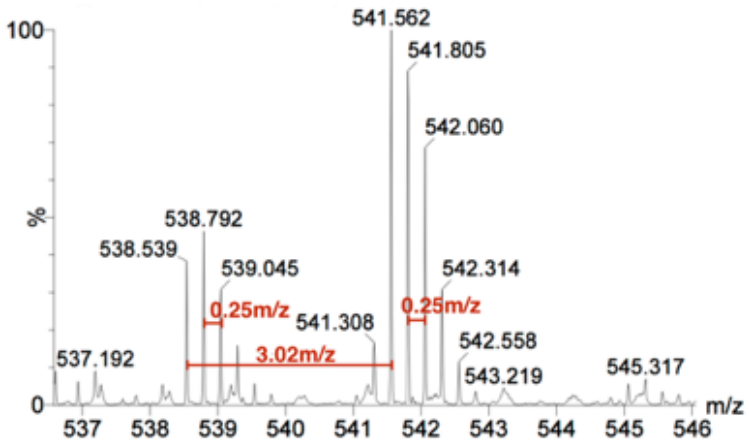
encoded reporter ions used in the labelling process are generated through fragmentation at low energies. In order to provide efficient fragmentation of both the reporter ions and the peptides a method was developed in which 50% of scan time is conducted at a low static energy and for the remaining time the energy returns to the defined peptide ramp.[115]

Finally, to assess if any benefit from the iTRAQ adaptation was due to energy range or to the temporal nature in which it is employed, a "Wide" energy ramp was also tested. This ramp incorporated the minimal and maximal energy values across all other ramps.

3.3 Results and Discussion

3.3.1 Validation of Score Threshold

(a) CCTKPESER-LSQKFPK, LD Score = 20, Charge +4, True identification.



(b) LVTDLKVHKECCHGDLLECADDR-EKVLASSAR, LD Score =17, Charge +5, False identification.

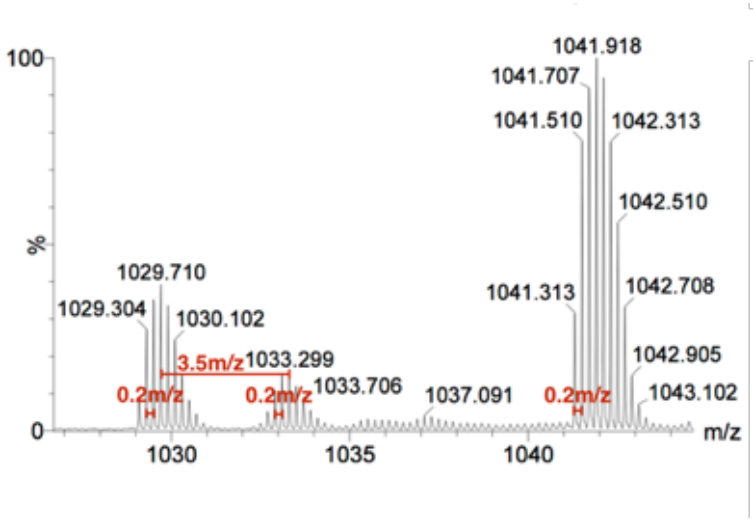


Figure 3.3: Representation of cross-linked precursor validation from raw data collected from cross-linked BSA. A) Spectra for a true cross-linked precursor identified by xQuest. B) Spectra for a cross-link identified by xQuest which is an incorrect assignment.

xQuest returns all cross-link identifications that exceed a pre-score filter. In the original published analysis a linear discriminant score (LD Score) threshold of thirty was recommended

to determine whether a cross-link is a true positive identification.[59] This recommendation was later revised to sixteen for tryptic digests. Leitner et al. [61] also demonstrated that the score threshold was dependent on the enzyme used during digestion.

To assess whether the LD Score threshold of sixteen was suitable for QToF datasets cross-link identifications were subjected to manual validation. In each case the m/z and retention time of the identified cross-link was used to isolate the species from the raw data file in Mass Lynx v4.1. The precursor scan is then inspected for features unique to cross-linked species. An example of such validation can be seen in Figure 3.3. Due to the use of pairs of deuterated cross-linker, cross-linked precursors will present with two approximate Gaussian distributions with a mass shift dependent on the cross-linker used. These represent the heavy and light versions of the cross-linker connected to the same two peptides in the cross-link. As such it is possible to verify the presence, but not sequence, of a cross-link reported by xQuest based upon the mass difference between the two distributions and the charge state. The charge state in turn can be determined by the mass shift between two isotopic peaks within the distribution.

The spectrum in Figure 3.3a was generated by a cross-link scoring 20.87. The spectra reveals two near-Gaussian distributions of peaks. The mass difference between the peaks is 0.25 m/z, in agreement with the +4 charge state reported by xQuest. In addition, the mass difference between both monoisotopic peaks is consistent with a mass shift of 3 Da. That is, the mass shift of the cross-linker (12 Da) divided by the charge state. In this case evaluation of the precursor spectrum reveals a cross-linked peptide.

The spectrum for the cross-link scoring 17.27 however, is reported by xQuest to have a monoisotopic mass of 1029.77 m/z (Figure 3.3b). Three approximate Gaussian distributions are found close to this m/z value, however no peak is found at 1029.77 m/z. This is likely due to absence of lock mass correction in the raw data. All three distributions posses a charge state of +5. The mass shifts between precursor distributions however, are higher than the expected mass difference between the cross-linker pairs. A mass shift of 2.4 Da between the monoisotopic peaks of the two distributions was expected. The observed mass shifts are 3.5 Da and 8.6 Da, for pairwise combinations from lower to higher respectively. The latter measurement has been removed from the figure for clarity. The absence of the correct mass

shift is indicative of a false positive cross-link identification.

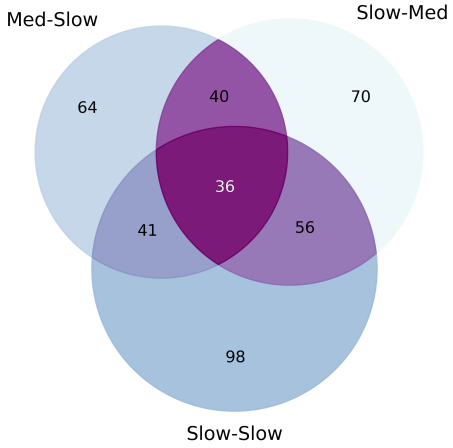
The validation process concluded that a threshold of sixteen often contained false positive identifications. Conversely, higher thresholds were found to miss many identifications that were confirmed. As discussed in Section 1.6.3 there is no way to establish a ground truth for cross-links, assignment of true positive and false positive identifications are subject to bias. Hence the creation of Receiver Operator Curves (ROC) was not carried out consequently a threshold of twenty was initially selected for further comparison of the cross-links in the triplicate dataset. Data were analysed by xQuest according to the workflow outlined in Section 2.4 (Materials and Methods). Although only minor adaptations were necessary, QToF datasets were not immediately compatible and required further conversion to mzXML format by MSConvert software with 32-bit encoding.

The choice of deisotoping algorithm was also found to have an impact on LD Scores. Protein Lynx Global Server v3.0.2 (PLGS, Waters Corp.) offers three levels of deisotoping that can be applied to both the precursor and the fragment ions. When processing the data using the "Fast" deisotoping algorithm in any combination, subsequent analysis by xQuest yielded no cross-link identifications (Table 3.3.1). Combinations of deisotoping featuring the Medium and Slow algorithms lead to differential cross-link identifications. Figure 3.4a reveals limited overlap between the Slow-Slow, Slow-Med and Med-Slow algorithms, with Slow on both precursor and fragment ion data achieving the optimal result. This combination was also found to generate higher xQuest scores for each cross-link identified (Figure 3.4b). The quality of the fragment ion data is considered by the xQuest scoring process and the Slow deisotoping algorithm considers lower intensity peaks.

Table 3.1: PLGS deisotoping algorithm test results. Deisotoping algorithms were combined in a pairwise manner at the precursor and fragment ion levels. The number of cross-links identified by xQuest and the highest score assigned to an identification is shown.

MS algorithm	**MSMS algorithm**	**Number of XLs**	**Highest Score**
Fast	Fast	0	0
Fast	Medium	0	0
Medium	Fast	0	0
Fast	Slow	0	0
Slow	Fast	0	0
Medium	Slow	73	41
Slow	Medium	79	44
Slow	Slow	108	50

(a) Overlap of cross-links identified for the three best performing PLGS deisotoping algorithm combinations.



(b) Custom heat map of identified cross-links from best three performing PLGS deisotoping algorithm combinations.

Figure 3.4: Cross-link overlap across all tested energy ramps. A) Venn diagram of cross-linked identifications by sequence. B) Heatmap showing cross-link ids by sequence and corresponding xQuest score assigned to each identification. Slow deisotoping of both precursor and fragment ion raw data provides the highest number of cross-link identifications with better xQuest scores.

3.3.2 Effects of Energy Ramps on Cross-link Identification Rates

Triplicate analysis of unique cross-link peptide pairs identified at different collision energies

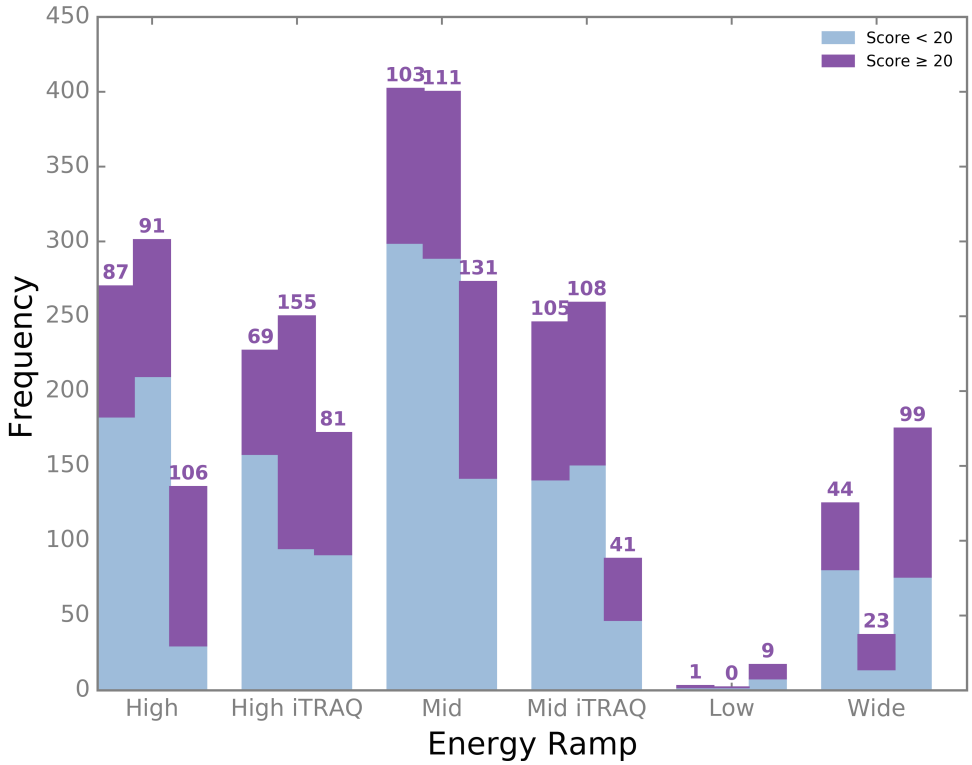


Figure 3.5: Comparison of xQuest scores for all identified unique BSA cross-link peptide pairs across six energy ramps. xQuest LD Scores are shown ≥ 20 (purple), ≤ 20 (lilac). Numbers above bars indicate a count of cross-links scoring ≥ 20.

Figure 3.5 shows the results of the triplicate analysis of all the tested energy ramps. All identified unique BSA cross-linked peptide pairs identified by xQuest are shown. Unique cross-links include those with linkages in the same absolute position in the peptides but with varying peptide lengths and/or modifications such as oxidised methionine.[57] Overall, reproducibility is good between the triplicate experiments of each ramp. The Mid ramp identifies the most reproducible number of high scoring cross-links: 103, 111, and 131 in the respective triplicate analysis.

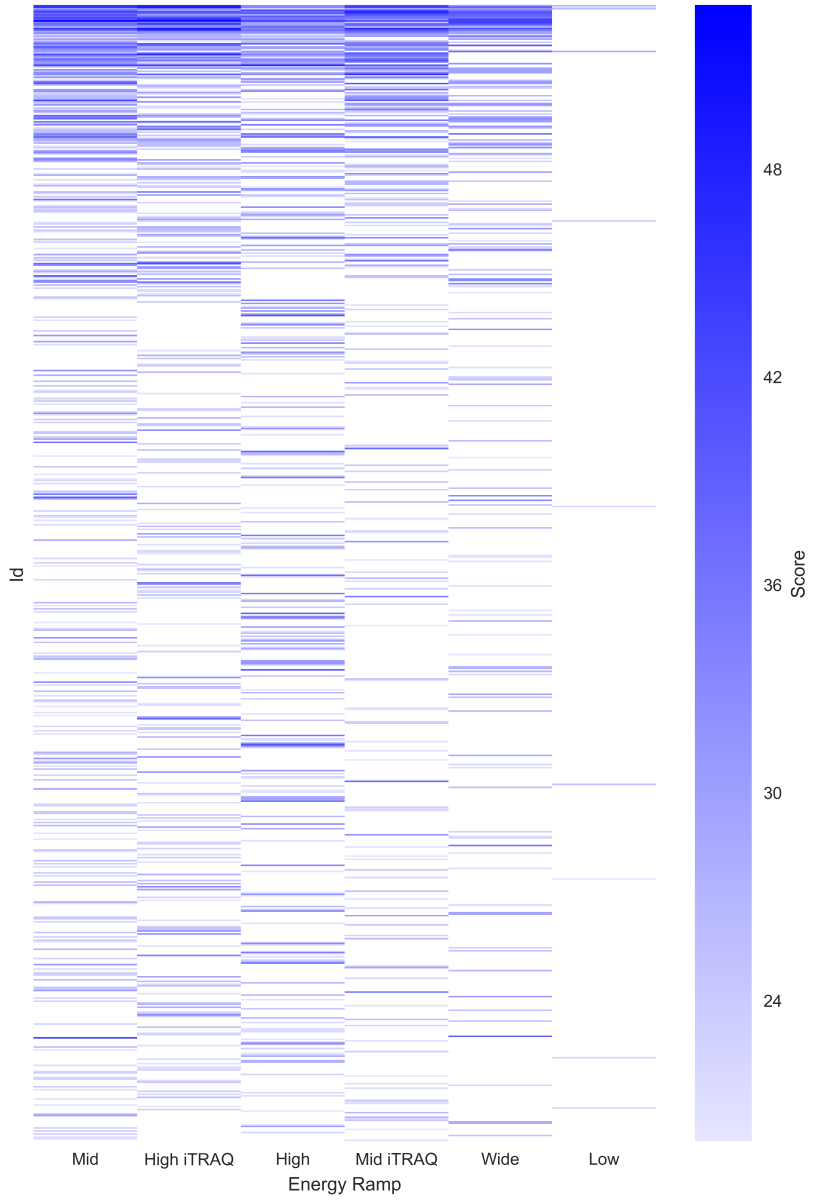


Figure 3.6: Cross-link overlap by sequence across each energy ramp tested. Each cross-link identified is coloured based the final score given by the xQuest software. These scores range from 20.0 to 52.79. Where no colourisation is displayed a cross-link has not been identified by that particular ramp. To aid interpretation ramps are arranged in descending order of the number of cross-link identifications and cross-links are displayed in descending order of number of ramps in which they can be found. Cross-links found in all ramps displayed at the top. Due to large divergence between cross-link identifications individual cross-link IDs have not been displayed.

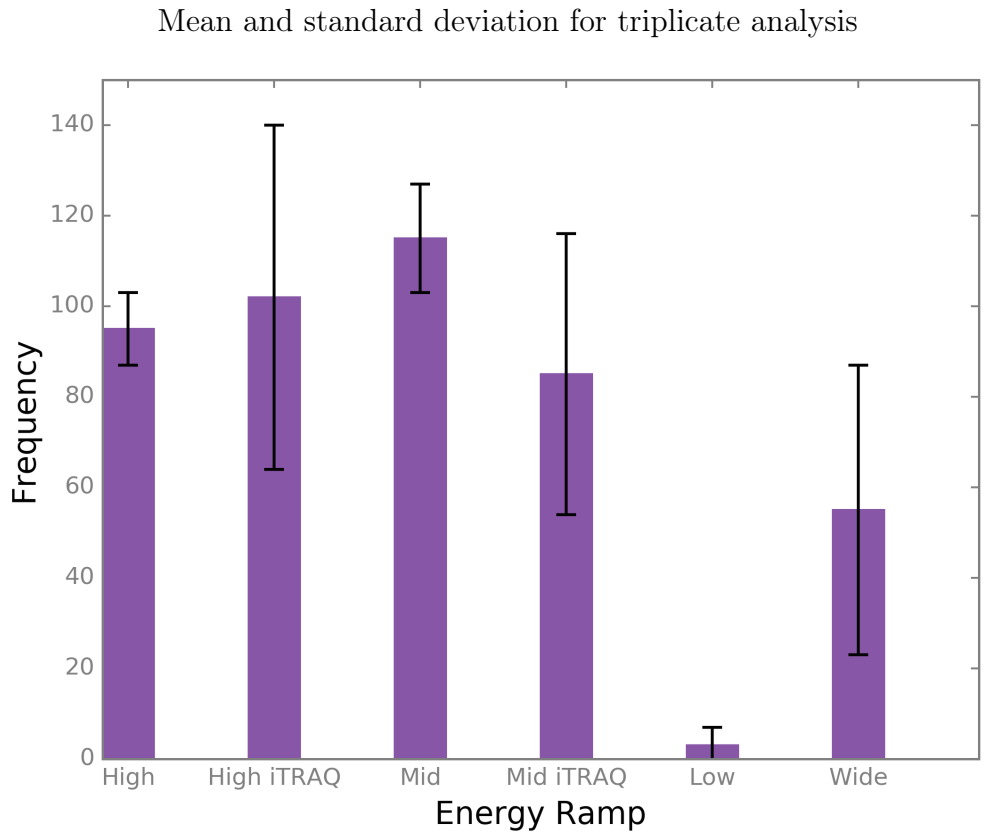


Figure 3.7: Mean and standard deviation for number of identified validated unique BSA cross-link peptide pairs in triplicate analysis of all 6 energy ramps. Cross-links have been validated as described in section 3.3.1. The mid energy ramp displays the greatest number of identified cross-links with the smallest variability across technical repeats.

Figure 3.7 shows the mean and standard deviation of the identified cross-links validated according to the method described in Section 3.3.1. Energy ramps that occupy the widest range of energies, the Wide and both iTRAQ adaptations, display the greatest variability in identification rates. The iTRAQ method devotes 50% of the scan time to a lower energy range whilst the Wide ramp encompasses the full extent of the energies tested. As these ramps sample a broader range of energies over the same scan time the time spent at each energy is reduced and only the most labile bonds can fragment. The selectivity of the energy range is therefore reduced leading to greater variability.

The High iTRAQ ramp also identifies a large number of cross-links across triplicate runs: 69, 155 and 81. Only 77 of the cross-links identified with the High iTRAQ ramp appear in the Mid ramp (Table 3.2). This motivates increased cross-link yield by applying different energies ranges for analysis. In this case an increase of up to 28% for unique cross-link identifications

was observed.

Table 3.2: Quantitative overlap of unique BSA cross-links identified in pairwise combinations of each energy ramp. The Mid and Mid iTRAQ combination yields the highest number of cross-links. Total count from intersection of the triplicate analysis of each ramp highlighted in yellow.

Ramp	**High**	**High iTRAQ**	**Mid**	**Mid iTRAQ**	**Low**	**Wide**
High	233	-	-	-	-	-
High iTRAQ	60	238	-	-	-	-
Mid	63	77	277	-	-	-
Mid iTRAQ	59	63	81	191	-	-
Low	4	4	4	3	8	-
Wide	40	47	60	44	4	143

3.3.3 Cross-link Validation by Solvent Accessible Surface Distance

The most frequently reported method of assessing cross-link validity published in the literature is calculation of cross-link length.[87,61,114,16,99,29] In almost all publications this length is calculated as the Euclidean distance between two residues. As reported by Bullock et al. [13] this distance is often inaccurate as the path of the cross-link is blocked by the protein structure. Cross-links must circumvent the protein rather than pass through it, as such the Solvent Accessible Surface Distance (SASD) is a more accurate measure. SASD is the shortest possible path between two amino acids that does not penetrate the surface of the protein. In order to evaluate the lengths of the cross-links identified by xQuest the distance distributions of the identifications with a LD Score above twenty were plotted (Figure 3.8). Through evaluation of published results Bullock et al. [13] determined the most accurate SASD for the DSS/BS3 cross-linker was 33 Å. This length has been used to distinguish between the approved (blue) and violated (red) cross-links in each of the energy ramps. In most publica-

tions all identified cross-links are generally plotted in such a diagram. As the purpose of this work was to evaluate the segregative power of a scoring threshold the subset of cross-links that score above twenty have be used in this analysis.

SASD cross-link distances were determine by Jwalk v1.3.[13] It can be seen that the discriminative power of the score threshold is poor (Figure 3.8). Greater than 50% of the cross-links identified by each of the ramps are between amino acids that are more than 33 Å apart (Table 3.3). Although it is likely that some of these identifications will be false positives such a large proportion of violated cross-link distances is unexpected. The distance cut-off of 33 Å determined in Bullock et al. [13] was calculated on datasets where only cross-links between lysine residues had been considered. This distance was calculated between the carbon α atoms of both residues. As previously discussed in Section 1.6.1 NHS esters conjugate not only lysines, but also serine, threonine and tyrosine residues. Cross-links between these residues will possess a varying range of lengths. This indicates that this cut-off may not be a reliable discriminator.

Table 3.3: Quantity of accepted and violated cross-links from the intersection of the triplicate dataset from each ramp. Cross-links were evaluated on Solvent Accessible Surface Distance. Violations represent cross-links with a carbon α to carbon α distance greater than 33 Å.

Experiment	**Total SASD Violations**	**Total SASD Approved**
High	141	92
HighiTRAQ	137	101
Mid	165	112
MidiTRAQ	119	72
Low	5	3
Wide	87	56

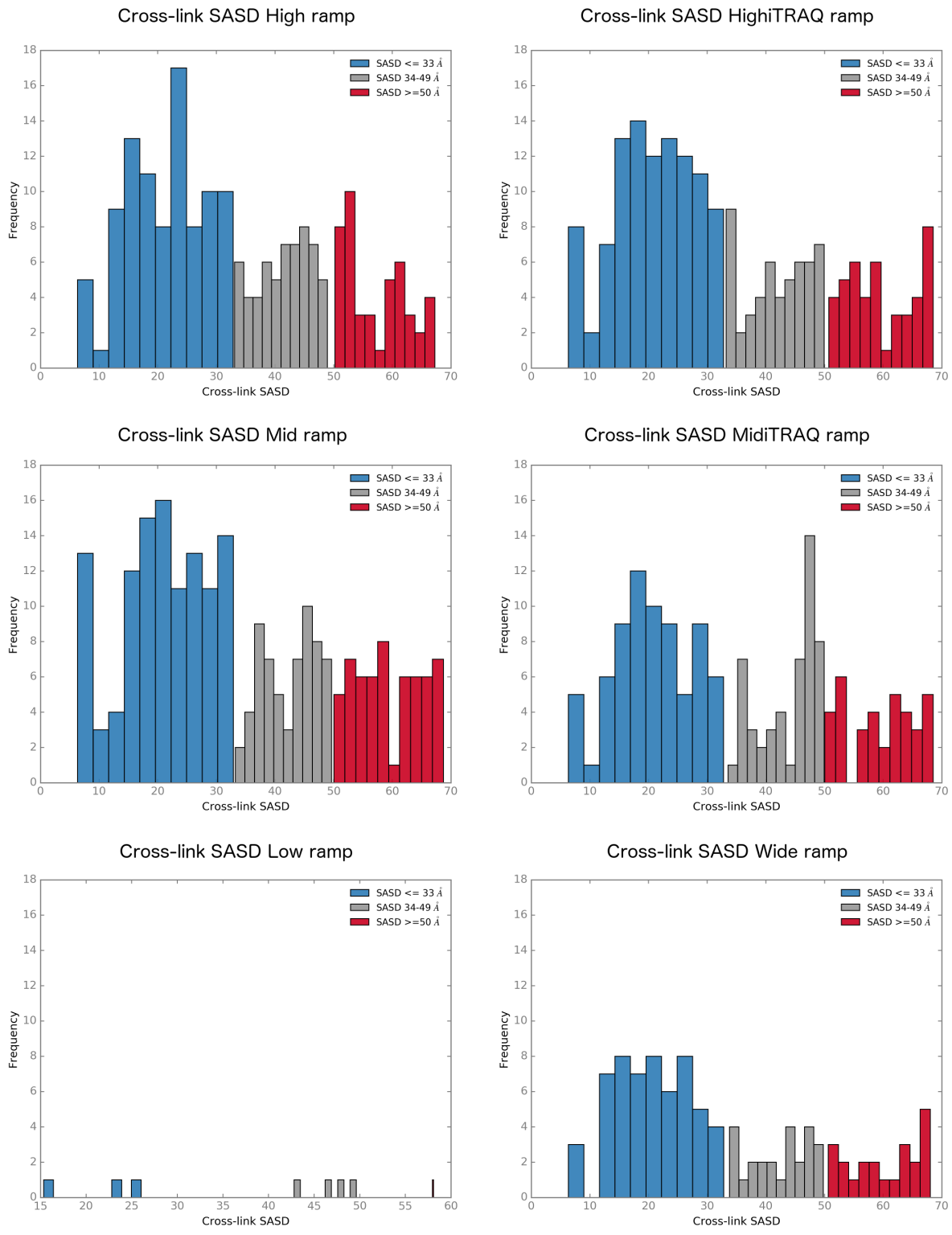


Figure 3.8: Distance distributions of cross-link length for each energy ramp. JWalk has been used to calculate cross-link SASD using BSA model PDB 4f5s. Cross-links longer than 50 Å are definite violations and shown in red. Cross-links scoring above 20 in each of the energy ramps have been used. Many violations of cross-link distance can be seen. Score threshold is not enough to determine quality of a cross-link validity.

3.3.4 Effect of Energy Ramps on Fragmentation Patterns

During MS/MS experiments cross-link fragmentation provides two ion types: cross-linked fragment ions which contain amino acids from both peptides joined by the cross-linker and linear fragment ions which are generated from either the α or β peptide which do not contain the linker (Figure 3.9). Both ion classes are necessary to assigned a candidate cross-link to an MS/MS spectra. Linear fragment ions help to confirm peptide sequences. Cross-linked fragment ions are essential to confirm that the peptides are cross-linked and not the result of an isobaric species.[47]

Figure 3.9: Schematic representation of fragment ions generated from a cross-link. Linear ions are shown in green, cross-linked fragment ions are shown in red.

In order to assess the distribution of these ion types, cross-link spectra were inspected. The spectrum for one of three cross-links identified by all energy ramps is shown in Figure 3.10. Linear fragment ions are shown in blue and cross-linked ions in red. The spectrum for the cross-link analysed by the High energy ramp displays no cross-linked fragment ions. In addition, the spectra generated with the Low and Wide energy have poor sequence coverage. Despite this, xQuest scores both these cross-link identifications favourably at 26 and 30 respectively. The base peak in the Wide and MidiTRAQ spectra corresponds to the precursor. The large energy range sampled by the Wide ramp and the temporal nature of the iTRAQ adaptation likely prevent efficient fragmentation. The Mid energy ramp offers the highest sequence coverage for this cross-link.

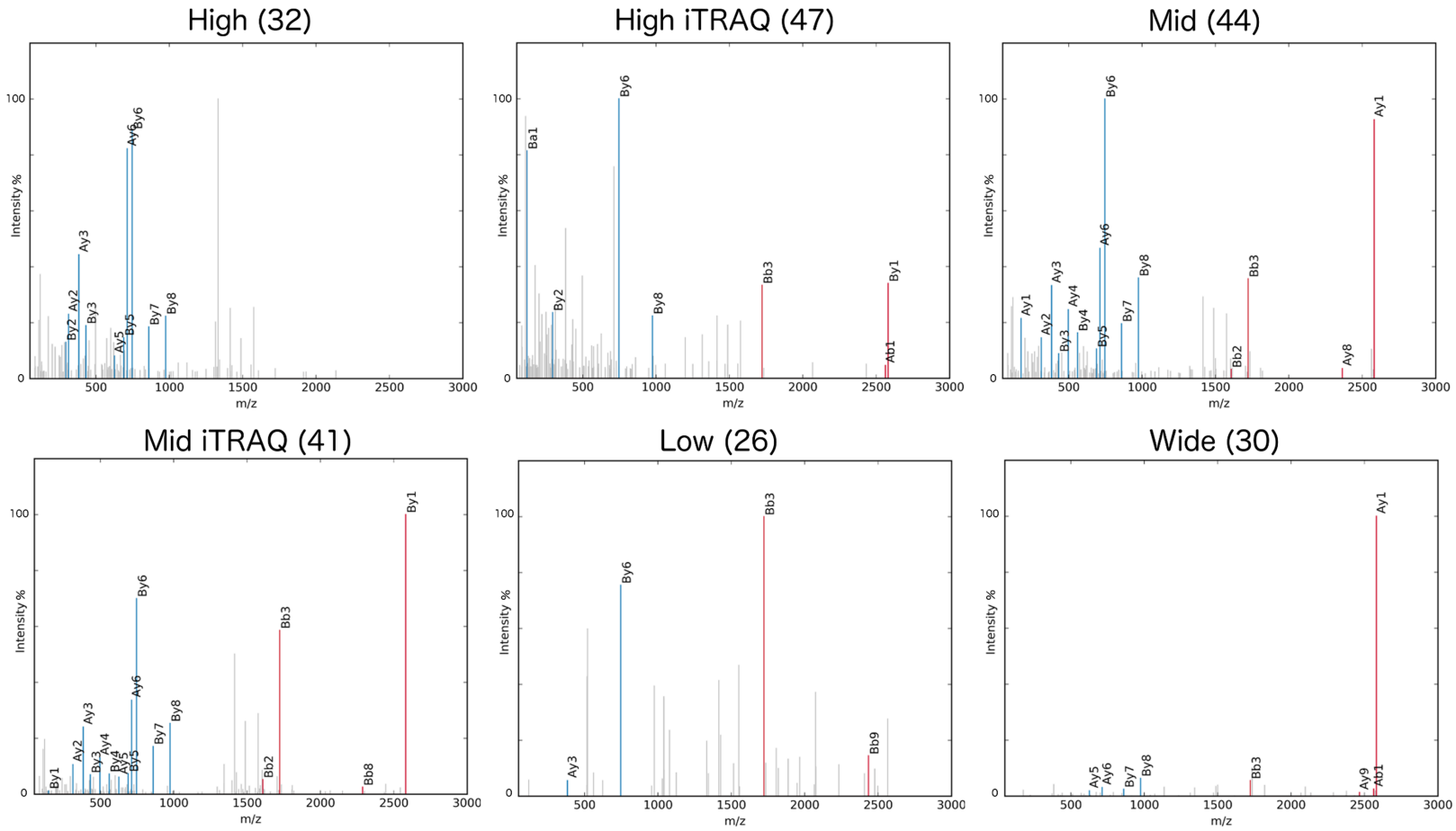


Figure 3.10: Example spectra for cross-link ID DTHKSEIAHR-FKDLGEEHFK (a4, b2). Cross-linked fragment ions shown in red, linear fragment ions in blue. Grey peaks represent unannotated peaks in the spectra. xQuest scores are shown in brackets. Spectra were created using AnnotateXL.py explained in more detail in Chapter 6.

As part of the xQuest scoring function the two types of fragment ions are assessed on the correlation between the observed and theoretical fragment ion spectra. This is performed separately for each type of ion. The results of these correlations are reported as two subscores: "XCorrx" for cross-linked fragment ions and "XCorrb" for linear fragment ions. To further assess the effects of the energy ramps on the identified cross-links the results of these scores were analysed in more detail. Both the mode and the mean of the XCorrx scores for the higher energy ramps (High and High iTRAQ) are considerably lower than for the XCorrb scores. In fact the most frequently reported XCorrx score for the High energy ramp was negative (Table 3.4). As the final score is a weighted sum of each subscore this negative result causes additional penalisation to the final score. In contrast the Low energy ramp shows the opposite affect, with a higher mode and mean reported for the XCorrb score.

Table 3.4: Descriptive statistics for the XCorrx and XCorrb subscore of the triplicate intersection across each energy ramp. XCorrx represents the cross-linked fragment ions and XCorrb represents the linear ions.

Ramp	**XCorrx Mean**	**XCorrx Mode**	**XCorrx S.D.**	**XCorrb Mean**	**XCorrb Mode**	**XCorrb S.D.**
High	0.02	-0.02	0.06	0.35	0.33	0.15
High iTRAQ	0.10	0.07	0.07	0.25	0.18	0.14
Mid	0.14	0.09	0.09	0.21	0.2	0.12
Mid iTRAQ	0.17	0.16	0.10	0.22	0.16	0.14
Low	0.26	0.24	0.11	0.10	0.05	0.05
Wide	0.19	0.12	0.11	0.21	0.19	0.12

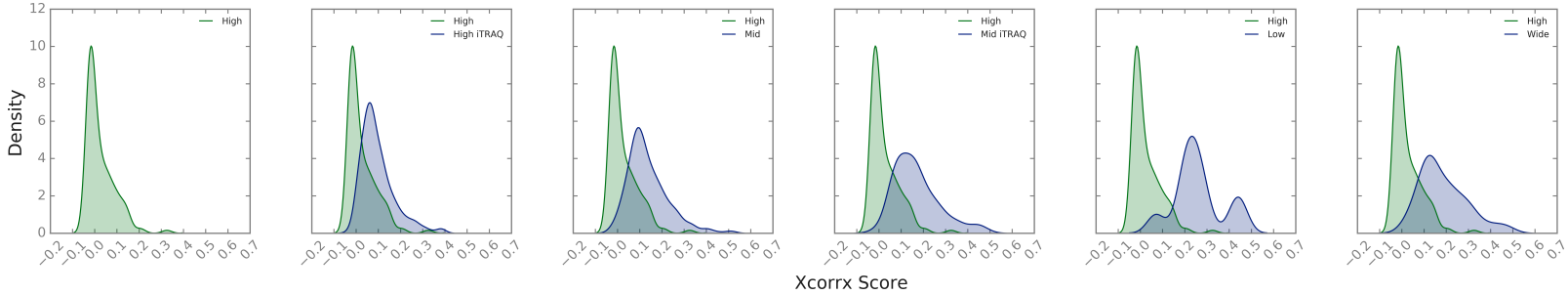
As the identified cross-links represent a sample of the wider population a kernel density estimation (KDE) was carried out on the XCorrx and XCorrb subscores. KDE estimates the underlying probability distribution for a sample provided that the identifications are independent and identically distributed. A full description of the technique is given in Appendix C.

Figure 3.11 shows a comparison of the estimated distributions for both the linear fragment ion correlation score (XCorrb, Figure 3.11b) and the cross-linked fragment ion correlation score (XCorrx, Figure 3.11a). As the High energy ramp was observed to give negative correlation scores for cross-linked ions the data is presented as a pairwise comparison of the

High energy ramp against all other ramps.

Figure 3.11b reveals contradictory performance of the Low and High energy ramps. Improved correlation scores for linear fragment ions are observed at high energies. At low energies these ions appear to fragment less efficiently. The opposing phenomenon is seen in Figure 3.11a. At low energies cross-linked fragment ions are preserved resulting in higher correlation scores. At higher energies the correlation score is lower due to the absence of cross-linked ions from the MS/MS spectra. In order to preserve both types of fragment ion during CID a medium energy ramp is required. This is reflected by the higher number of cross-link identifications in the Mid energy ramp.

(a) Pairwise comparison of XCorrx (cross-linked fragment ions) score KDE for all energy ramps. High energy ramp is shown in green, other ramps shown in blue.



(b) Pairwise comparison of XCorrb score (linear fragment ions) KDE for all energy ramps. High energy ramp is shown in green, other ramps shown in blue.

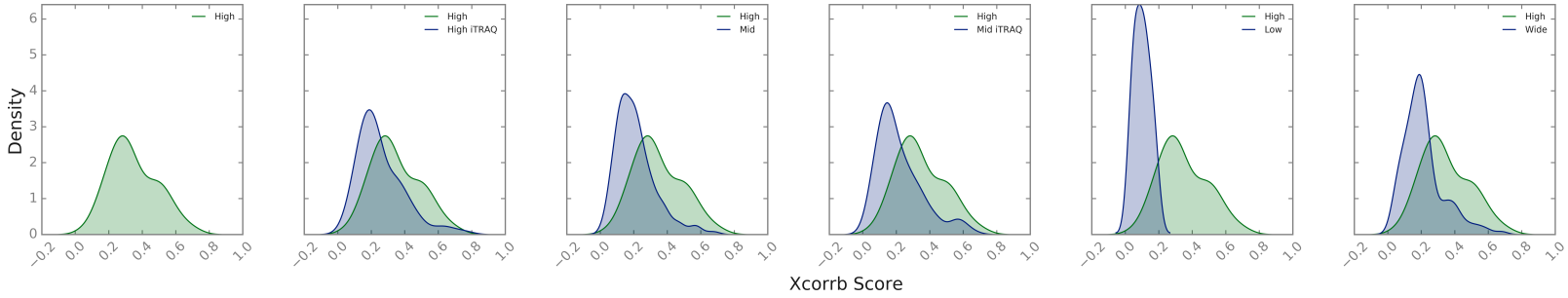


Figure 3.11: Comparisons of kernel density estimations for fragment ion correlations. High ramp shown in green, all others shown in blue. Ramps are compared in the following order in both A) and B): High with; High, HighiTRAQ, Mid, MidiTRAQ, Low and Wide. Cross-linked peaks receive higher scores at lower energies whereas linear peaks received higher scores at higher energies. To ensure the presence of both peak types a Mid range ramp is more optimal.

Ions Omitted from xQuest Searches

To further explore the reasons for the varying score distributions the fragmentation patterns associated with the tested ramps were evaluated. As negative XCorrx scores were observed with the High energy ramp this data was analysed for evidence of cross-linker fragmentation.

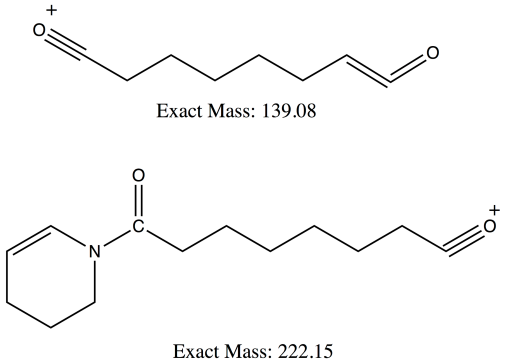


Figure 3.12: Representation of BS3/DSS diagnostic ions as previously identified by Iglesias, Santos, and Gozzo [48]. Figures produced using ChemDraw Professional 16.0. Masses for diagnostic ions have been calculated and conform to those previously published in literature.[48] The diagnostic ion at mass 222.15 represents the tetrahydropyridine modification to a lysine side chain.

In a previous analysis of synthetic cross-linked peptides BS3/DSS was shown to fragment at the cross-link amide bond[48] generating diagnostic fragment ions (Figure 3.12). These ions have also been identified during HCD analysis of cross-linked protein digests. Trnka et al. [111] identified the tetrahydropyridine modification in 71% of cross-link spectra. The results of our analysis find this modification in 52% of cross-link spectra that were analysed by the High energy ramp (Table 3.5). The poor correlation scores for cross-link peaks identified at this energy are likely due to fragmentation of the BS3 amide bond. It should be noted, however, that fragmentation of the BS3/DSS linker is not as readily observed when conducting CID in an Orbitrap. In their comprehensive study of cross-linked peptide CID behaviour Giese, Fischer, and Rappsilber [33] found this to be rare, appearing in only 10% of cross-linked peptide spectra.

Table 3.5: Percentage of cross-linker ions that have been modified as shown in Figure 3.12 which are present in spectra containing cross-links. MGF files were searched using an in-house script to calculate the percentage of diagnostic ion masses.

m/z	**High Ramp** %	**Mid Ramp** %	**Low Ramp** %
139.1	11	4	0
222.1	52	15	3

In addition to the diagnostic BS3 ions, two further types of fragment ions were found to be ignored by the xQuest algorithms. xQuest does not consider ions generated through fragmentation of both peptides in a cross-link. It assumes a singular fragmentation event per peptide. That is, fragmentation of cross-linked ions will only occur on the alpha or beta peptide but not on both.

Furthermore, immonium ions are also not considered. The range of the LIT is limited to 200-2000 Da, whereas a ToF or an Orbitrap has a 50-5000 Da range. This includes masses of immonium ions created during fragmentation (Figure 3.13). These ions are diagnostic of the presence of specific amino acids in a peptide and are often used in *de novo* peptide sequencing.

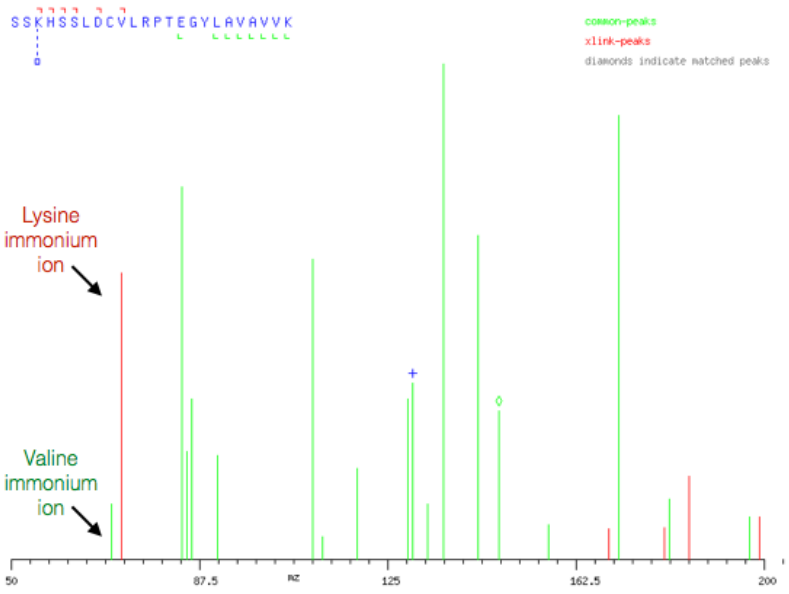


Figure 3.13: Mis-identification of cross-linked peaks in QToF data by xQuest at the range not considered in an LIT (50-200 m/z). Spectra for monolink SSKHSSLDCVLRPTEGYLAVAVVK is shown. Valine immonium ion and lysine ($-NH_3$) immonium ion are indicated. Due to the 12 Da mass shift between the peaks the lysine ($-NH_3$) immonium ion has been erroneous identified as a cross-linked peak by xQuest software.

During analysis of MS/MS spectra these ions were found to hinder the xQuest scoring algorithms. Figure 3.13 shows an example of this for a monolink identified in the Mid energy analysis. The immonium ions for lysine (84 Da) and valine (72 Da) are present in the MS/MS spectra for this monolink. The mass shift between these ions is 12 Da, and as such xQuest has identified the peak at 72 Da as a linear peak (green) and the peak at 84 Da as an erroneous cross-linked peak (red). xQuest is unable to annotate the peaks since the algorithms do not expect the presence of immonium ions. This leads to unannotated peaks in the spectra and a reduced overall score. Although scores were slightly penalised when the mass range was increased cross-link identification rates were unaffected. Modification of the xQuest mass range is therefore not recommended when searching QToF data.

3.4 Conclusion and Further Work

The application of QToF mass spectrometers to most fields of MS research is commonplace however, they have yet to be widely applied to cross-linking analysis. We have demonstrated

that despite differences in operational parameters and MS data, existing cross-linking software can be incorporated into a workflow with only minor modification. We have identified an optimal collision energy range that preserves the unique ion species generated from cross-links during fragmentation. This work reveals that by conducting CID with multiple energy ramps, cross-link yield can be improved by up to 28%.

In addition, xQuest also does not consider cross-linked ions where fragmentation events have occurred on both the peptides. Although these problems should not be overlooked, they have not prevented high numbers of cross-link identifications in this analysis. The scoring algorithms have been designed to assess spectral quality however, they are limited by the types of ions they are designed to account for. Validation of the results is time consuming and the duration increases with cross-link yield. In addition, employment of a simple score threshold serves only as a guide. It would be beneficial to include all fragment ion types in the evaluation of each cross-link identification and to assess the sequence coverage.

The method presented in this work for the analysis of cross-linked peptides on a QToF motivates the addition of ion mobility separation (IMS) to a cross-linking experiment. This gas phase fractionation is conducted online and may lead to a reduction in sample preparation requirements by removing the need for prior enrichment. Furthermore, recent developments in IMS methods may allow the DDA process to select more highly charged species for MS/MS analysis.[34]

Chapter 4

Ion Mobility Enhanced Data Dependent Acquisition for the Analysis of Cross-linked Peptides

4.1 Introduction

Ion mobility is a gas phase separation technique which segregates ions based upon their size, shape and charge. The technique is widely used in biological mass spectrometry to separate mixtures of isomers, polymers and chiral compounds. It has applications in the fields of proteomics, metabolomics and glycomics.[53] There are many different implementations of this separation technique.[21,90,95,15,38] All make use of an inert buffer gas and an electromagnetic field in order to separate the ions.

The earliest application of this method relies on the diffusion of ions through a drift tube filled with inert gas. Since this was the pioneering implementation it is often referred to as Ion Mobility Separation (IMS). As discussed in Section 1.4, the progress of the ions through the cell is solely by directed diffusion. This relationship permits the calculation of a mobility constant for an ion based on its Collision Cross Section (CCS) (Section 1.4 Equation 1.6).

The mobility constant of an ion is also affected by its charge state. Taraszka, Counterman, and Clemmer [103] used IMS to separate a complex mixture of fourteen tryptically digested proteins. Analysis revealed that ion mobility segregated the peptides into different charge

state families. This "separation according to charge" has been shown to persist across multiple implementations of ion mobility separation methods. Pringle et al. [83] showed that when using Travelling Wave Ion Mobility Separation (TWIMS) a tryptic digest of four proteins also displays segregation between peptide charges states. In this case the largest separation was found to exist between singly and multiply charged species.

In addition to charge state separation ions from different classes of biomolecule can also be separated. Fenn et al. [28] calculated the CCS for a series of nucleotides, peptides, lipids and carbohydrates as a function of their m/z. Differential mobility was observed for each biological species with a corresponding m/z. Further analysis of a mixture of these components using IMS revealed that lipids had the greatest degree of separation, emerging last from the mobility cell. Nucleotides emerged first with carbohydrates and peptides following respectively. This separation was attributed to intra-molecular folding forces. Deceased hydrophobic in lipids reduces their packing density compared to peptides, nucleotides and carbohydrates in the gas phase.

TWIMS has also been used to study biological separation.[104] It should be noted however, that when using TWIMS CCS values cannot be directly calculated from the mobility constant of an ion. As described in Section 1.4.1 (Figure 1.6a), TWIMS uses an alternating radio frequency pulse in a concentric pattern across the stacked ring ion guide, the energy available to the ions is greater than the energy provided through interactions with the buffer gas. As such mobility is not proportional to the field strength. The ion's drift time is therefore more accurately referred to as an 'arrival time'. In order to generate CCS values when using TWIMS calibration of the instrument is required.[37]

A further method for generating CCS values for peptides was proposed by Thalassinos et al. [104] and requires the creation of a calibration curve from a series of known peptide CCS values. The CCS values for a series of unmodified peptides and peptides containing phosphorylated residues was the estimated. Highly phosphorylated peptides displayed smaller CCS values than peptides that had an equal m/z. The smaller mobility constant of these phospho-peptides was also attributed to intramolecular folding. As the phosphate modification is charged it is likely that the peptide backbone will exhibit folding around the phosphate group creating a more compact structure.

As cross-links contain two peptides joined by a linker they differ from their linear counterparts in both charge, shape and size. Given that ion mobility has been used successfully to separate different biological species and different charge state families of the same species, the possibility of cross-link separation from unmodified peptides has been explored in this work. Optimisation of the mobility separation was carried out, followed by an analysis of the effects of the technique on cross-link identification rates. Consideration has been given to the effects of enrichment and of sample complexity on mobility separation as well as the addition of pusher synchronisation to the experimental parameters.

4.2 Materials and Methods

4.2.1 Preparation of uncross-linked BSA

In order to compare the mobilities of linear and cross-linked BSA peptides a sample of uncross-linked BSA was prepared as follows: 0.3 mg/ml BSA (A7030, Sigma-Aldrich) was prepared in 20mM HEPES @ pH 7.6 and evaporated to dryness. To ensure the comparison to the cross-linked sample was accurate, following in-solution tryptic digest, the sample was cleaned using solid phase extraction and fractionated using a MicroAkta as previously described in Section 2.2.

4.2.2 IM-DDA Experimental Design

Samples were analysed using the gradient and settings described in Section 2.3 with the best performing collision energy identified in Section 3.3.2: LM 10-20 eV HM 30-60 eV. Data were acquired using Data Dependent Acquisition (DDA) with the addition of mobility separation (IM-DDA). The following parameters were optimised by using the mobility pattern of the calibrant, Glufibrinopeptide-B so that the first twenty and last ten drift time bins contained minimal ion signal intensity. A wave height of 40 V and IMS gas flow of 90 mL/min were used. As presented in Section 4.3.1 wave velocity was optimised to be 500m/s. In contrast to the previously reported DDA method Collision Induced Dissociation (CID) fragmentation was performed in the transfer following mobility separation.

4.2.3 Extraction of Mobility Time of Linear and Cross-linked BSA

When generating an IM-DDA method with fragmentation in the transfer the arrival times for precursor ions are reported in both the MGF and XML output that is generated by Protein Lynx Global Server v3.0.2 (PLGS) following raw file processing. Full processing parameters are given in the Section 2.4 and have been kept constant across all experiments.

For the unmodified BSA a search was performed in PLGS to identify the peptides. The following search parameters were used: fixed modifications of carbamidomethylation of cysteine, variable modifications of oxidised methionine, two missed cleavages were considered, with a 5 ppm and 10 ppm tolerance for precursor and fragment ion matching respectively. A fasta database containing BSA with common contaminant proteins was used in the search.

Cross-link searches were performed as described in Section 2.4. For the identified cross-links xQuest reports the retention time, molecular weight and charge state of the precursor ion. It is possible to match these characteristics back to the MGF or XML file in order to extract drift times for the cross-linked ions.

4.3 Results and Discussion

4.3.1 Optimisation of Mobility Parameters for Cross-linked Peptides

In order to generate optimal separation of the charge state families of a cross-linked BSA tryptic digest, a range of IMS wave velocities was sampled. During these tests the IMS wave height and the gas pressure were both maintained and wave velocity varied to 300 m/s, 400m/s, 500m/s and 600m/s respectively. A plot of m/z as a function of arrival time (hereafter referred to as the mobility plot) for each of the wave velocities is shown in Figure 4.1.

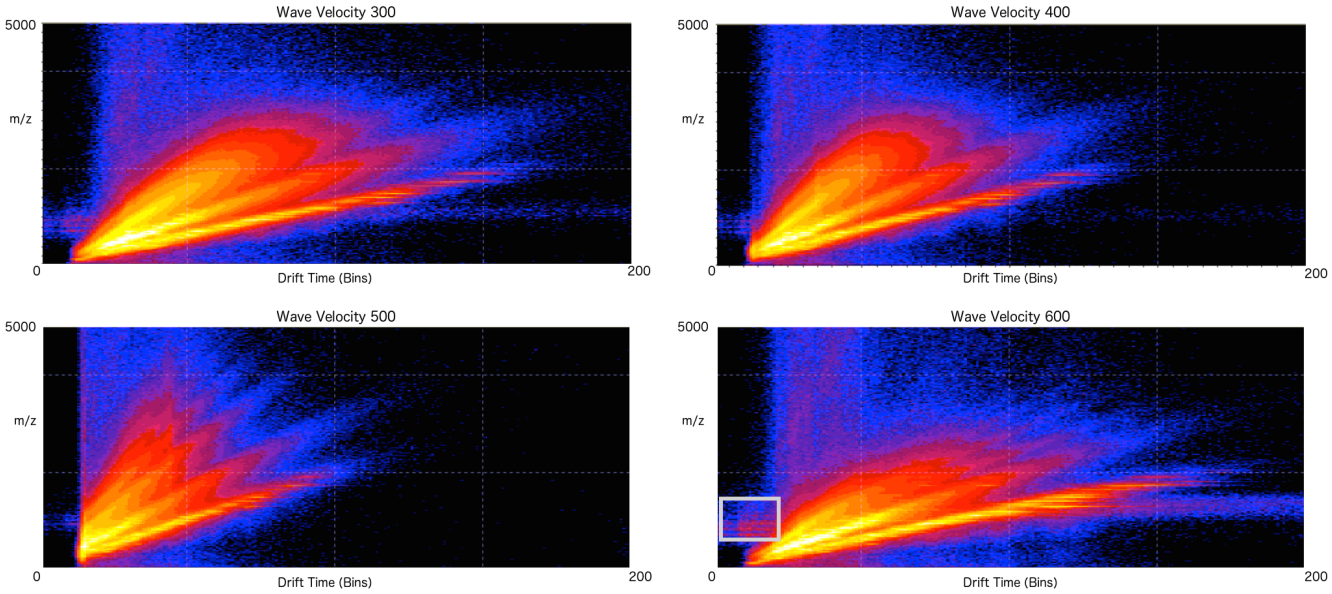


Figure 4.1: Wave velocity optimisation. Velocities of 300, 400, 500 and 650 m/s were tested for optimal mobility separation. Mobility plot generated from survey scan using DriftScope v2.8. Intensity Threshold values Min=30% and Max=100% counts using a logarithmic map intensity scale. Grey box indicates roll over. Wave velocity of 500 m/s provides optimal separation of the precursor charge states.

It can be seen that as the wave velocity increases the separation between each charge state family also increases until the wave velocity reaches 600m/s, at which point the ions are observed to coalesce. As the wave velocity is defined as the distance between the stacked ring ion guides divided by the length of time in which the travelling pulse is applied, lowering the wave velocity increases the length of time a pair of ring electrodes are exposed to the pulse. This gives better separation of charge states as the ions are exposed to the field for longer. As cross-linked peptides have a higher charge state than linear peptides better separation may enable higher number of cross-links to be identified.

In addition, for a wave velocity of 600m/s, horizontal streaking due to rollover is observed in the region between 700 and 1000 m/z (Grey box Figure 4.1. This is caused when a packet of ions fails to reach the pusher before a new packet of ions is released. Ions of the same m/z are observed in multiple drift time bins. Thus, at this wave velocity, the time taken to traverse the mobility cell by ions of this m/z is too long.[72]

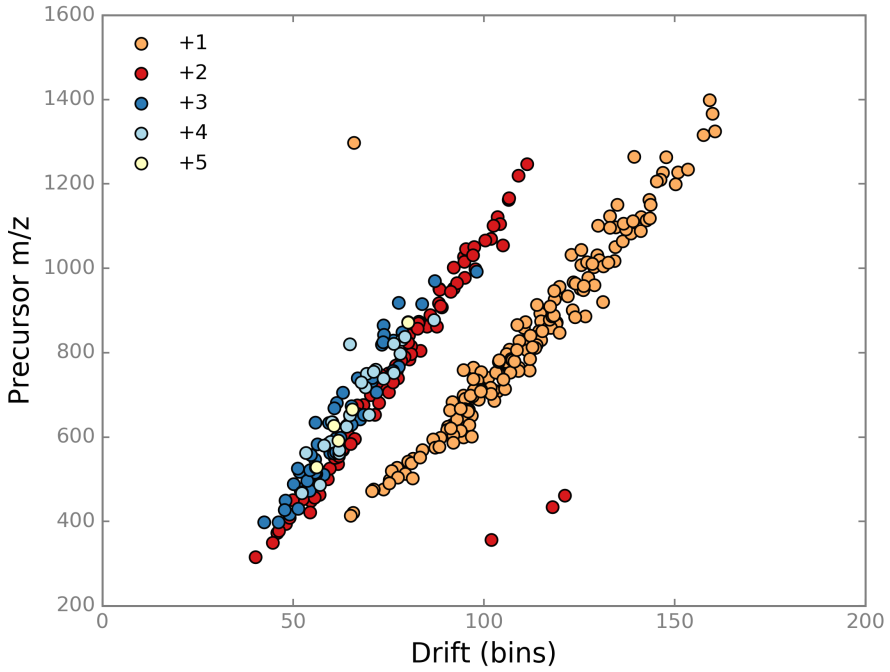
Finally, it can be seen that the first 12 bins for each wave velocity are empty. In order to make maximal use of the mobility separation the empty bins were trimmed in subsequent

methods. A wave velocity of 500m/s provides the optimal separation of charge state families and was employed in all further experiments.

4.3.2 Mobility of Cross-linked and Linear BSA Peptides

Using the optimised mobility settings a sample of cross-linked and unmodified BSA was analysed. Mobility times for the identified cross-links and peptides were extracted. Figure 4.2a shows the mobility plot of the identified peptides for the unmodified BSA. There is clear separation between the singly and multiply charged species. In addition to the expected charge states for tryptic peptides ranging from +1 to +3, a number of +4 and +5 charge state also exist. A full description of the peptides with charges states above +3 can be found in Appendix D. With the exception of the peptide at 580 m/z, arrival time bin 58, VLIAFSQYLQQCPFDEHVK these peptides all represent at least one missed cleavage event.

a) Mobility of unmodified BSA peptides as a function of m/z and charge state



b) Overlap of cross-linked and linear BSA peptide mobility as a function of m/z

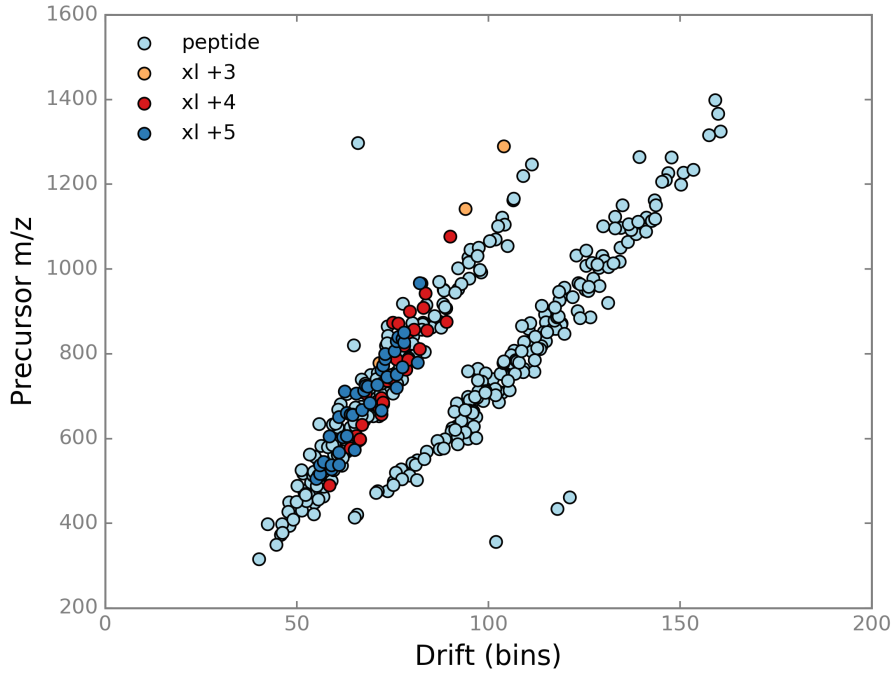


Figure 4.2: Mobility plots for linear and cross-linked peptides a) Comparison of linear BSA peptide mobility across all charge states. Separation of singly charged peptides in mobility space is clear. b) Comparison of cross-linked and linear BSA peptide mobility. For clarity, charge state is differentiated for cross-linked peptides only. Cross-linked peptides overlap with linear peptides in mobility space.

Figure 4.2b shows the mobility plot for the identified BSA cross-links overlaid with the linear peptide data from Figure 4.2a (light blue). The majority of identified cross-links have a charge state of +4 or +5. Despite the optimisation of the wave velocity parameters both the cross-links and the higher charged linear peptides are not separated in different regions of the mobility plot. A limited amount of separation however, is observed above 1000 *m/z*. Closer inspection of signal intensity from the regions of higher charge state in the mobility plot reveal poor signal intensity. In good agreement with Figure 4.2 the majority of the signal intensity identifiable from the analysis is located in the region thought to be occupied by the +2 charge state on the mobility plot.

The acquisition method for this technique is fundamentally DDA. During analysis cross-linked peptides are frequently much lower in intensity compared to their linear counterparts. It is therefore possible that their signal is masked during the precursor ion selection process. Giles et al. [34] developed a method to enhance the signal-to-noise ratio of low abundance ions by using ion mobility to remove the signal generated by singly charged ions. By removing the singly charge ions from our dataset an increase in sensitivity of the precursor section process can be achieved enabling a more targeted DDA analysis.

4.3.3 Enhancement of IM-DDA with the Application of Charge Stripping

To develop a targeted DDA approach a rule file must first be generated from the sample of interest. Following analysis of the cross-linked BSA sample the region of the mobility plot which excludes the singly charged ions was selected in DriftScope v2.8 (Figure 4.3). This was then exported to create a rule file: a tab delimited text file which provides a start and end *m/z* for all arrival time bins. This file is used to synchronise the pusher to pulse ions within the *m/z* range into the ToF analyser. This is performed by applying a temporal delay to the pusher pulse based on the transit time of the ion through the mobility cell.

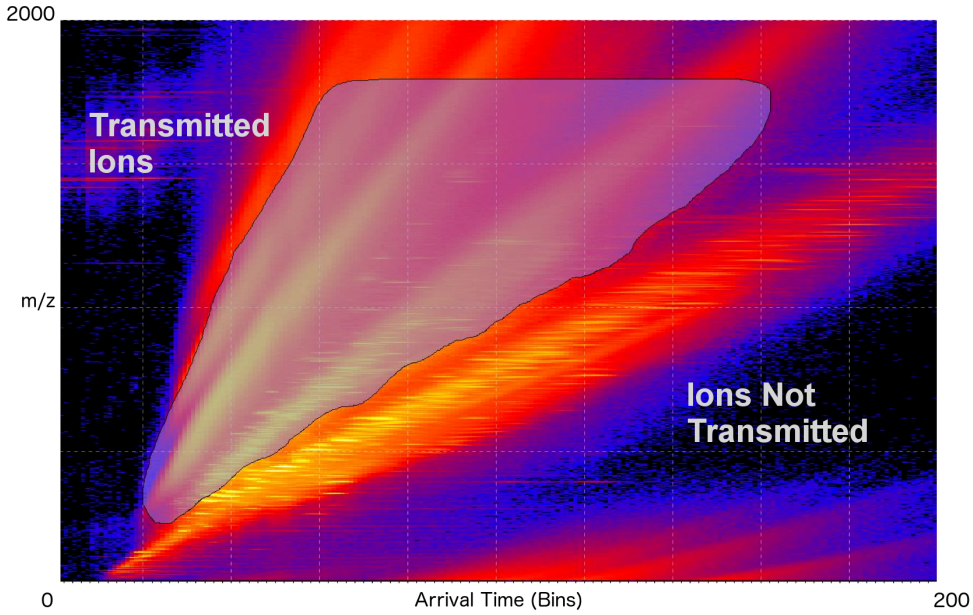


Figure 4.3: Discriminating ion transmission. Ions on the left in highlighted area are transmitted by pusher synchronisation. Ions on the right are not transmitted to the detector. Rule file for IM-DDA pusher synchronisation was generated using DriftScope v2.8 (Waters Corp.).

4.3.4 Comparison of Identified Cross-links across Mobility and Non-Mobility DDA

Figure 4.4 shows a comparison of the unique cross-link peptide pairs identified by analysis with the DDA, IM-DDA and IM-DDA with charge stripping methods as discussed above. The DDA method identifies a higher number of cross-links in all three of the triplicate analysis however, 50% of the identified cross-links had a xQuest lines discriminant score below twenty and as discussed in Section 3.3.2, are likely false positive identifications. The experiments performed using both ion mobility methods show a great reduction in the number of false positives, with fewer than forty cross-link identifications.

Number of unique BSA cross-link peptide pairs identified by xQuest following IM-DDA analysis

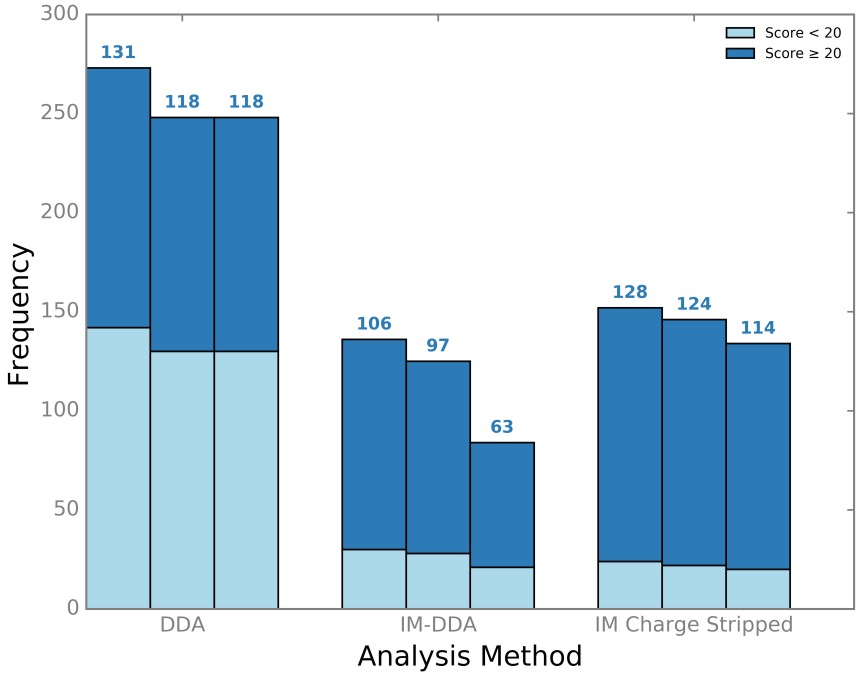


Figure 4.4: Comparison of triplicate analysis for all identified BSA cross-links for each analysis method. Bars display count of unique cross-links identified, this includes cross-links with the same absolute residue position but with sequence modifications such as oxidised methionine residues. Cross-links with xQuest scores above 20 shown in dark blue, those with scores below 20 shown in light blue.

As ion mobility separates species with similar m/z and retention times in each scan incorporation of the method may allow overlapping species to be isolated for fragmentation. If more than one species exists in a packet of ions following isolation by the quadrupole, fragmentation of the mix of precursors will yield product ions that cannot be assigned to a single precursor. This would cause mis-annotation of product ion peaks and affect the overall xQuest linear discriminant score assigned to a cross-link. Analysis of the MS data failed to isolate any such species however, the extra degree of separation provided by the addition ion mobility separation would prevent this overlap.

Mean and standard deviation for validated cross-links identified from cross-linked BSA.

Experiment	Mean	Std
DDA	122	6
IM-DDA	89	18
IM-DDA Charge Stripped	122	5

Table 4.3.4 displays the mean number of validated cross-links for each analysis method. These cross-links were validated according to the method in Section 3.3.1. The IM-DDA method without the addition of charge stripping has the fewest identifications, however the standard deviation is larger and close to the range of the number identified in both the other methods. This variability may be due to the presence of high intensity singly charged species interfering with the DDA precursor selection process. The charge stripped mobility method compares favourably to the DDA method with similar numbers of validated cross-link identifications. However, a further increase in cross-link identifications was expected.

4.3.5 Effects of SEC on IM-DDA analysis of cross-linked peptides

The absence of an observed increase in cross-link identification rates for both mobility methods may be due to the enrichment of the sample by size exclusion chromatography (SEC). As the sample has already been subjected to prior separation by size, increases in signal due to ion mobility separation may be masked. In order to ascertain the extent of these effects a sample of un-fractionated cross-linked BSA digest was analysed using the IM-DDA Charge Stripped method.

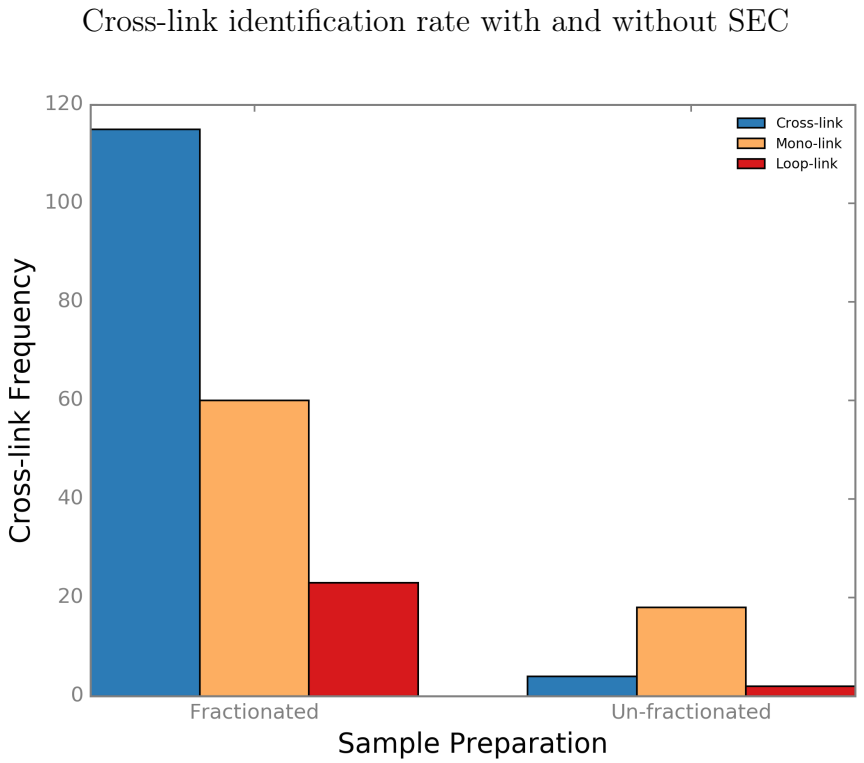


Figure 4.5: Effects of SEC on validated cross-link type and identification rate. Graph shows number of intra molecular cross-links, mono-links and loop-links identified by xQuest analysis of data collected with and without size exclusion chromatography separation prior to mass spectrometric analysis.

Figure 4.5 shows the identification rates for all types of validated cross-linked peptide, with a xQuest linear discriminant score above twenty, identified from each of the samples. The sample enriched prior to analysis reveals a significantly higher identification rate for every type of cross-linked product. It is therefore unlikely that SEC separation is responsible for a reduction in cross-link identification rates when using ion mobility separation in the experimental workflow. The results of this experiment indicates that enrichment is still necessary when analysing cross-linked samples with the addition ion mobility separation.

4.3.6 Effects of Sample Complexity on Cross-link Identification Rates with Mobility and Non-Mobility Methods

When ion mobility separation is employed in the analysis of complex mixtures, different biological species segregate readily. As such separation may become more effective as the

complexity of the sample increases. To examine the effects of sample complexity on ion mobility separation of cross-linked peptides nine protein monomers were cross-linked in isolation. The resulting samples were digested and purified. The peptides were then mixed and enriched with SEC. Full methods for the preparation of the 9 mix sample are given in Section 2.2.

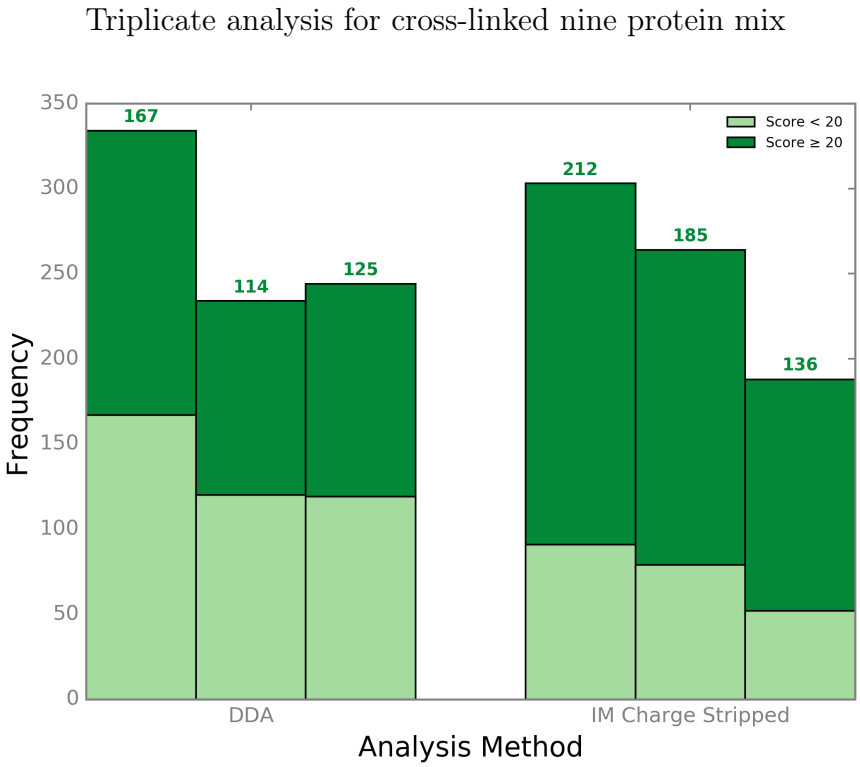


Figure 4.6: Comparison of number of cross-link identification rates for all unique cross-linked peptide pairs found for the nine protein mix samples analysed with DDA and IM-DDA charged stripped methods. Cross-links with xQuest scores above 20 shown in dark green those with scores less than 20 shown in light green.

The samples were analysed using the standard DDA method and the IM-DDA charged stripped method. Figure 4.6 shows the full range of unique cross-link identifications. In good agreement with the analysis of cross-linked BSA ion mobility analysis of the nine protein mix has fewer low scoring cross-links than the DDA method. As previously discussed in Section 4.3.4 this is most likely due to improved isolation of precursors leading to a more accurate annotation of the fragment ions.

Mean and standard deviation for validated cross-links identified from the nine protein mix

Experiment	Mean	Std
DDA	135	23
IM-DDA Charge Stripped	178	31

The mean number of cross-link identifications for each experiment is shown in Table 4.3.6. When compared to the DDA analysis, a greater increase in the number of validated cross-links scoring over twenty is seen for the IM-DDA with charge stripping analysis of the nine protein mix. The increase is however, within the experimental error given by the standard deviation. As such the addition of ion mobility to the experiment has yet to provide a noticeable improvement in cross-link identification rates.

4.3.7 Reduction in singly charged precursors

The principle of the IM-DDA Charge Stripped method is to increase the sensitivity of the DDA analysis by removing more intense singly charged ions that may prevent precursor selection of cross-linked ions. This is achieved through the synchronisation of the pusher lens in the ToF (Section 4.3.3). One possible explanation for the observed lack of increase in cross-link identification rate may be due to inadequate removal of the singly charged ions. In order to determine whether the charge stripping was accurately applied during the experiment the proportion of different charge state species from the MGF file was compared for both DDA and IM-DDA charge stripped experiments for the fractions in the nine protein mix sample (Figure 4.7).

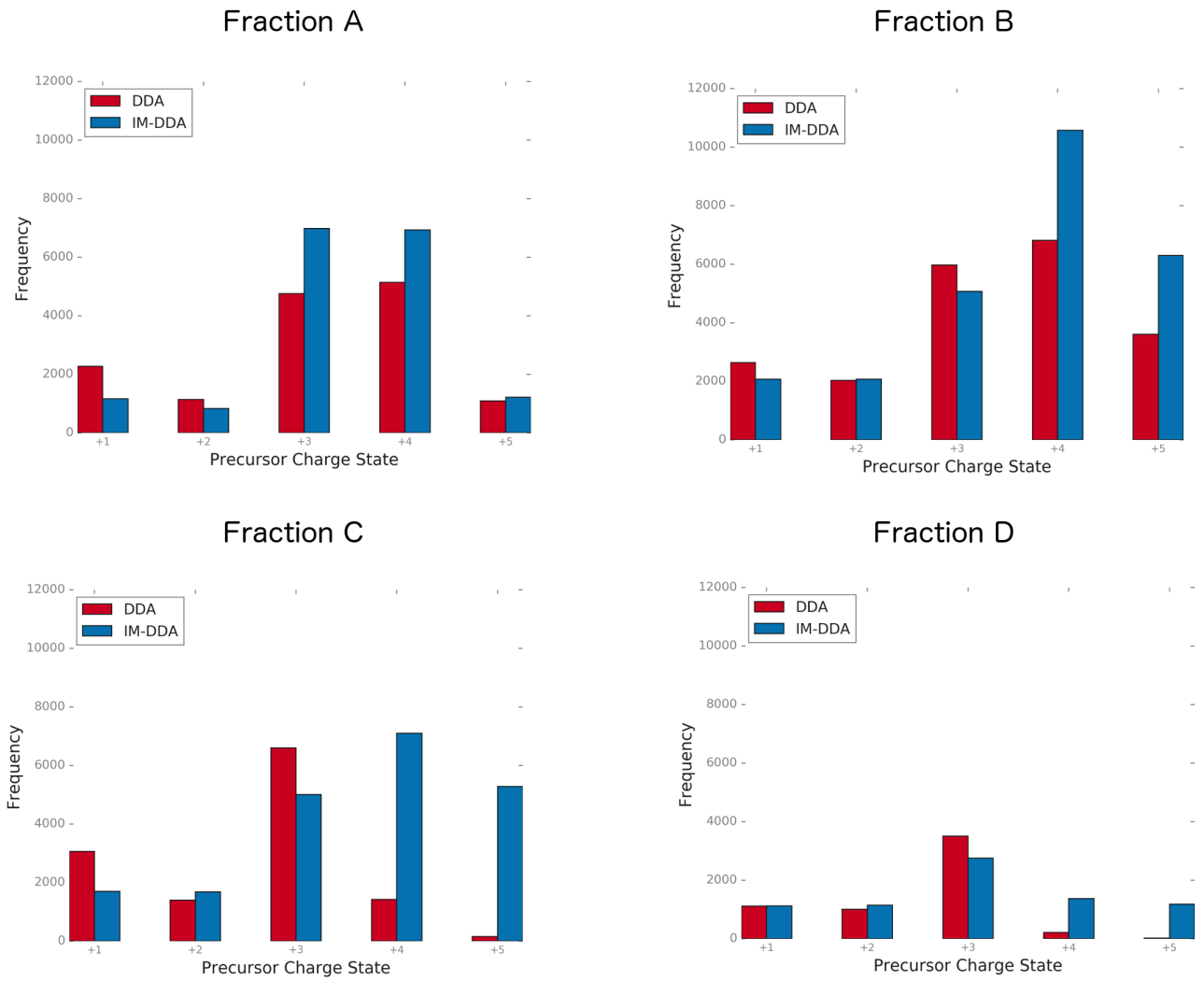


Figure 4.7: Charge state distribution for cross-linked BSA SEC fractions generated from MGF files. Precursors with charge states ranging from +1 to +5. No precursors with charge states above +5 were identified. Results from DDA analysis shown in red, IM-DDA analysis shown in blue. SEC fractions have been plotted separately. In most fractions the IM-DDA method identifies an increased number of higher charge state precursor ions.

Across all fractions analysed with the IM-DDA Charge Stripped method a decrease in the number of singly charged ions is observed. The only exception to this is fraction D where the number of singly charged ions recorded during the experiment is equal for both IM-DDA charged stripped and the DDA method. Despite the observed reductions, the total number of singly charged ions that have been analysed in each experiment remains high: with fraction B containing almost 2000, fractions A and D approximately 1000 and fraction C containing 1800 singly charged precursors. A larger reduction in the total count of these ions would be

expected if the pusher was correctly synchronised to the arrival time of higher state ions at the entrance to the analyser.

In contrast to this observation there is also an increase in the number of higher charge ions for the IM-DDA charge stripped method compared to the DDA method. This is observed to varying extents in all fractions and suggests that the pulse from the pusher lens is correctly synchronised. Figure 4.2b in Section 4.3.2 reveals that uncross-linked peptides with any charge state above +1 overlap. It also indicates that linear peptides are higher in abundance than cross-linked peptides. By removing the singly charged ions, cross-link ions are not directly targeted. The method simply encourages higher charge state precursor selection. The observed increase in higher charge states in the MGF files may correspond to a mixture of cross-linked and linear precursors with charge states above +3.

Presently the method does not appear to adequately remove the singly charged ions. This may explain the lack of increased cross-link identification rate despite an increase in isolation of higher charge state precursors. A deeper evaluation of the effects of the rule file inclusion to the method design is required to explain this contradiction. To date the sensitivity of the IM-DDA methods has not been greatly explored. It is possible that even with charge state selection methods, the signal from the highly charged linear peptides is still of sufficient magnitude to affect the precursor selection during the survey scan.

4.4 Conclusion and Further Work

Despite the previous success of ion mobility mass spectrometry in separating both charge state families in complex protein digests[96] and biological species in elaborate mixtures[28] separation of cross-linked precursor ions from linear peptides has not been demonstrated to any high degree. In a mobility plot cross-linked peptides within a range of 400 to 1000 m/z have been observed to overlap unmodified peptides (Section 4.3.1 Figure 4.2b). This may be due to the effects of intra-molecular folding forces.

Phospho-peptides for example, are separable from unmodified peptides using ion mobility and the degree of separation increases with the number of phosphate modifications.[104] This suggests that the peptides are able to wrap around the phosphate groups forming more

densely packed structures that move faster through the mobility cell. The cross-linked peptides generated in this study however, are joined by a long hydrophobic linker. Their structure is more likely to be branched with limited intramolecular folding within each linked peptide. As a result they may have similar CCS values to linear peptides. This could be confirmed with molecular dynamic simulations and by repeating cross-link analysis and arrival time extraction following generation of a CCS calibration curve.

Furthermore the charge state and size of a biological species have counter effects on TWIMS. Increasing charge accelerates the transit of a species through the mobility cell. Increasing size however, causes more collisions with buffer gas molecules slowing transit. For the cross-links studied in this work the effect does not appear to be consistent across the precursor mass range. In Figure 4.2b (Section 4.3.1) separation of cross-linked species from unmodified peptides at m/z values above 1000 is observed. Improved mobility separation may therefore be possible at higher m/z ranges. This could be achieved by using digestion enzymes that generate longer peptides within the cross-links. This effect would be dependent on the protein sequence. An alternative approach that could be applied without such dependence would be to perform a limited digest of the cross-linked sample. With further development to the experimental design it may be possible to isolate cross-linked species from linear peptides.

Chapter 5

High Definition Data Dependent Acquisition for the Analysis of Cross-linked Peptides

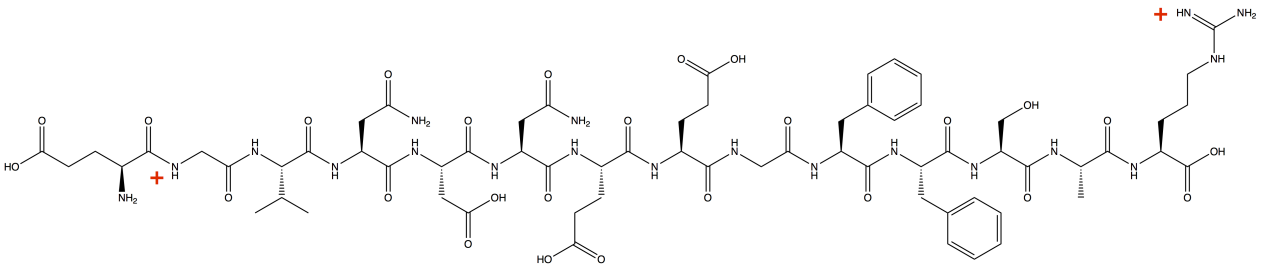
5.1 Introduction

The development of orthogonal acceleration Time of Flight (oa-ToF) mass analysers by Dawson and Guilhaus [23] enabled ToF analysers to be connected to continuous ion sources. Despite this enhancement, the sensitivity of an oa-ToF analyser is still limited by the duty cycle. The duty cycle of an instrument is defined as the ratio of the number of ions pushed into the analyser and the number of ions lost per spectral acquisition.[38] It is the proportion of time during which the analyser is usefully operated. As previously discussed in Section 1.3.2 the addition of a pusher lens above the entrance to the ToF delivers a potential that focuses a packet of ions into the analyser on an orthogonal trajectory from the continuous ion beam. Following the injection of each packet of ions the remaining beam passes directly through the pusher and not into the analyser. These ions are effectively lost, reducing the sensitivity of the instrument and reducing the duty cycle.

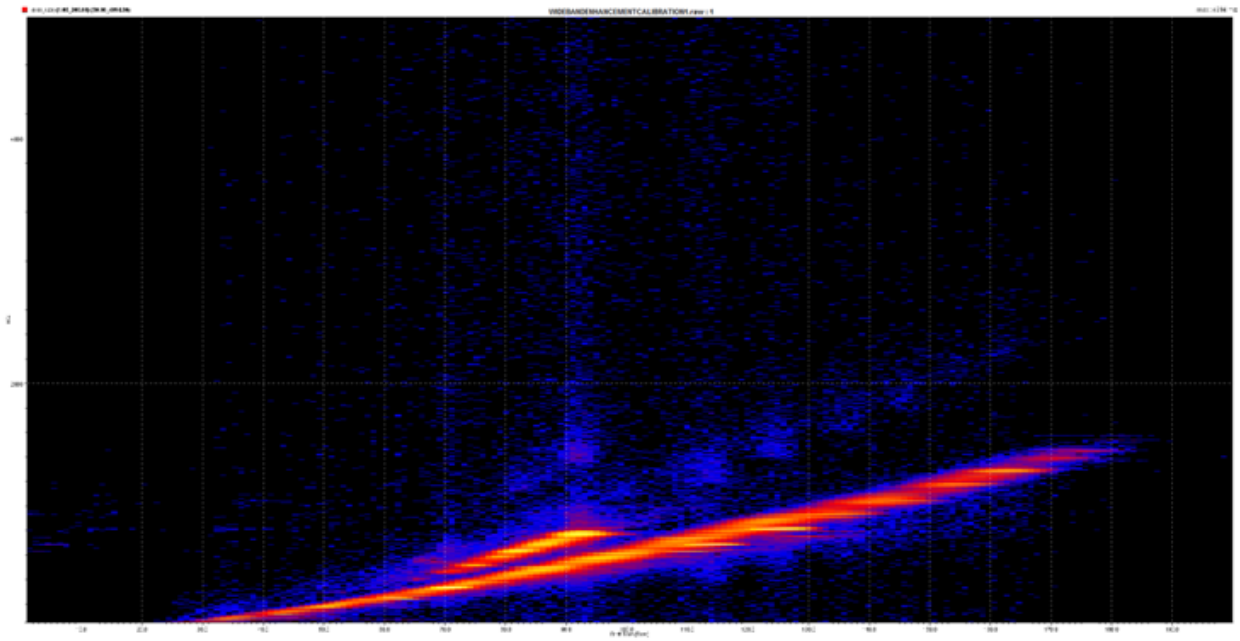
To improve sensitivity packets of ions can be stored in Travelling Wave Ion Guides (TWIG) following ion mobility separation. The pusher pulse can then be synchronised to the release of ions from the storage device, increasing the number of ions entering the anal-

yser.[38,44] In early studies of the technique Giles et al. [38] used a fixed delay of 29 μs on the pusher plate. This was shown to deliver an increase in duty cycle from 12 % to 100% for an ion of 684 m/z. As the time the ions take to exit the TWIG and reach the pusher is dependent on their m/z the fixed delay provided an enhancement in sensitivity across a fixed mass range, but with diminished sensitivity outside of these bounds.

To adapt the technique for a wider mass range the delay must be applied dynamically to the pusher. Helm et al. [44] achieved this for peptides by synchronising the pusher to the mobility of peptide fragment ions prior to their analysis in the ToF. In this method, known as High Definition Data Dependent Acquisition (HD-DDA), precursor ions are fragmented in the first of three TWIG devices in the Triwave: the trap. Gas phase separation of the product ions occurs in the mobility cell, separating ion species based on their size, shape and charge. By synchronising the pusher to the recorded mobility time of peptide fragment ions better sensitivity can be achieved across the full mass spectrum range.



(a) Structure of [Glu1]-Fibrinopeptide B (GFP) used as a calibrant for the arrival time of ions at the pusher lens in original HD-DDA method.[44] Positions of charge acceptance shown in red. Image created using ChemDraw Professional 16.0.



(b) Mobility pattern of product ions generated by fragmentation of Glufibrinopeptide-B (m/z 785.8427) at collision energy of 32eV in the Trap TWIG. Image generated using DriftScope v2.8. Intensity threshold values Min=30% and Max=100% counts using a logarithmic map intensity scale. Singly charged and doubly charged fragment ions can be seen to separate in mobility space.

Figure 5.1: Mobility pattern and structure of [Glu1]-Fibrinopeptide B calibrant.

In order to achieve this for a wider mass range dynamic synchronisation of the pusher was performed according to the mobility of Glufibrinopeptide-B (m/z 785.8427) fragment ions (Private Communication). When fragmented this doubly charged precursor generates predominantly singly charged fragment ions (Figure 5.1). In a method known as wideband enhancement, this mobility pattern was used to create a calibration file using DriftScope v2.8(Waters Corp.). The calibration file was then implemented into a Data Dependent Ac-

quistion (DDA) method to calculate the required delay for the pusher pulse based on the recorded exit time of an ion species from the mobility cell. Compared to a non-mobility DDA analysis of a HeLa cell digest, these adaptations led to an increase in MS/MS intensity. This increase generated an improved spectral acquisition rate of 60% leading to better scoring peptide identifications which were present in 25% more of the acquired spectra.[44]

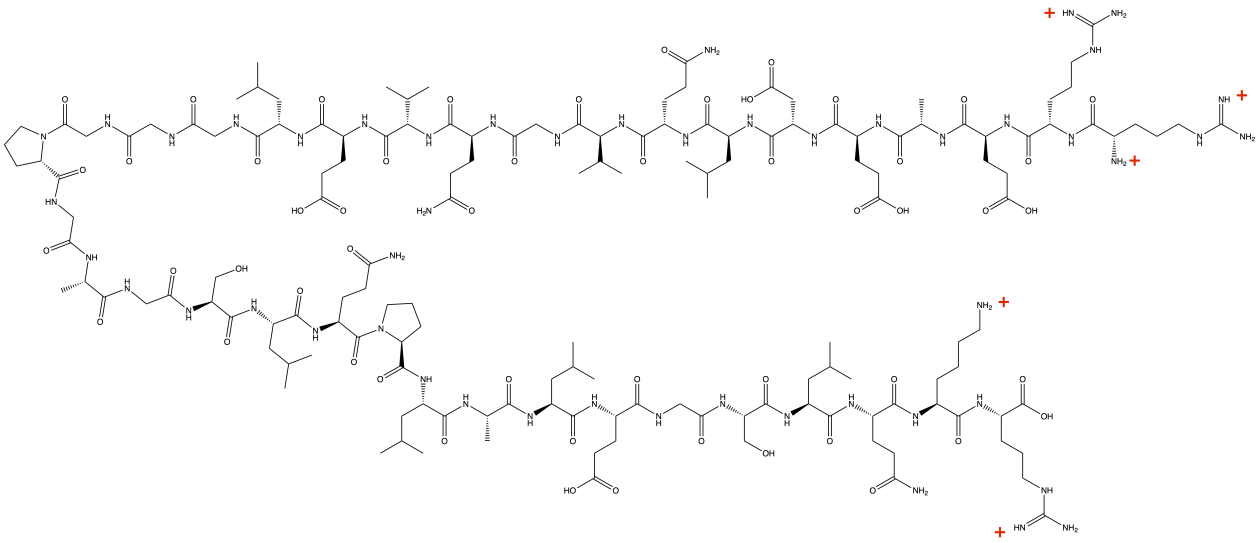
Cross-linked peptides are known to be underrepresented in the MS data. By increasing the duty cycle it may be possible to increase the number of cross-links identified by improving the sensitivity of fragment ions. An implementation of the HD-DDA method described above has been carried out in this work with adaptations to suit the physicochemical nature of cross-linked samples.

The use of the doubly charged Glufibrinopeptide-B precursor as a calibrant for wideband enhancement in the original HD-DDA method was an acceptable facsimile for a HeLa cell digest as tryptic peptides have a predicted mean charge state of +2.[5] Cross-linked peptides however, incorporate two covalently linked peptides. As such a charge state of +3 is most commonly is observed with a range that extends to +6.[33] The fragment ions generated during MS/MS analysis can range from +1 up to the charge carried by the precursor ion.

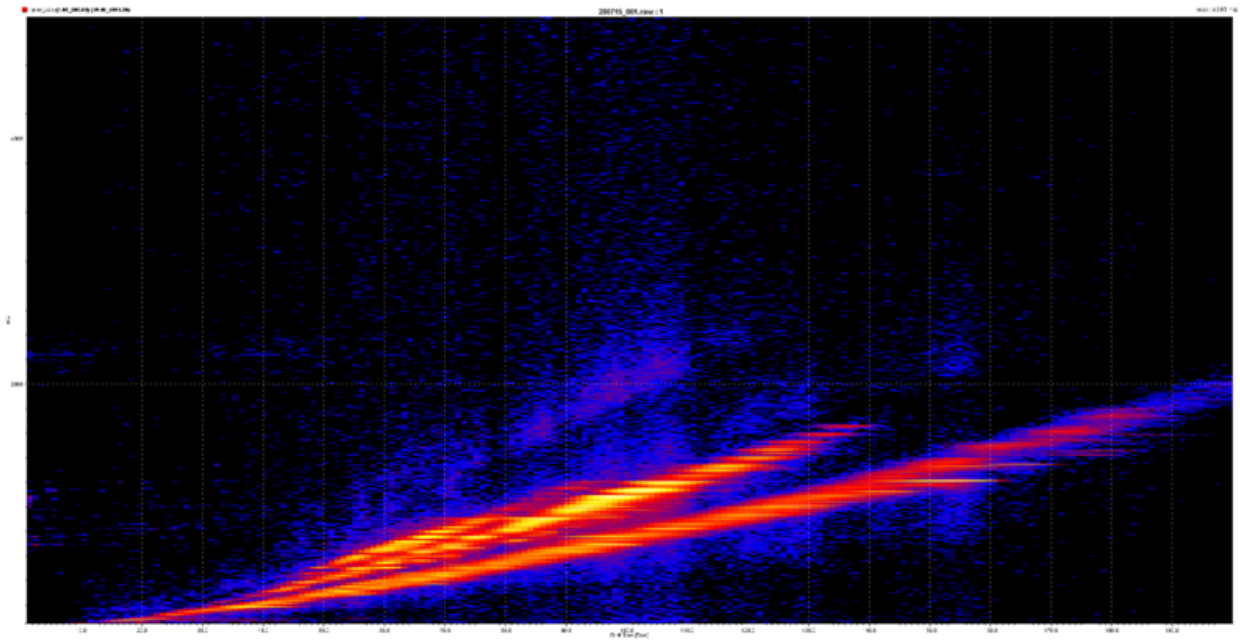
To synchronise the pusher to the arrival of fragment ions from cross-links, calibration files for all the observed charge state families must first be generated. These are then used individually to in separate experiments to synchronise the pusher pulse to the mobility of ions from each charge state family. In order to evaluate potential cross-link:spectrum matches the MS/MS spectra must be recombined. A computer program to merge the fragment ion peak lists for all precursors was written. The script uses tolerance parameters based on prior observations of precursor ions from a complex digestion. This procedure is explained in more detail in Section 5.2.2. In a first attempt calibration files for wideband enhancement were generated from the proinsulin C-peptide. Following evaluation of the dataset the sample itself was used as a calibrant. These files were incorporated into the experimental design and an analysis of their effects on the instrument duty cycle and cross-link identification rates was carried out.

5.2 Materials and Methods

5.2.1 Preparation and Analysis of Proinsulin C-peptide



(a) Structure of proinsuln C-peptide used as a calibrant for the arrival time of fragment ions at the pusher lens in the revised HD-DDA method presented here for the analysis of cross-linked peptides. Positions of charge acceptance shown in red.



(b) Mobility pattern of product ions generated by fragmentation of proinsulin C-peptide at collision energy of 41eV in the Trap TWIG. Image generated using DriftScope v2.8. Intensity threshold values Min=30% and Max=100% counts using a logarithmic map intensity scale. Single, double and triple charged fragment ions can be seen to separate in mobility space.

Figure 5.2: Mobility pattern and structure of proinsulin C-peptide.

The precursor to human insulin, proinsulin, consists of the insulin A and B chains separated by a 31 amino acid peptide known as the C peptide (Figure 5.2a). When analysed by Electrospray Ionisation mass spectrometry (ESI-MS) the C peptide precursor shows three distinct charge states; +1, +2 and +3, with +2 being the most abundant.[55]

Proinsulin C-peptide was purchased from Sigma Aldrich and made to a final concentration of 1 μM in 97% water 3% acetonitrile and 0.1% formic acid. To generate the calibration files necessary to synchronise the pusher proinsulin C-peptide was introduced into the mass spectrometer using a native source with the following settings: Cone voltage of 40 V, source temperature of 100°C. The +4 charge state of the precursor at 905.24 m/z was isolated with the quadrupole and trap collision energy was set to 41 eV. Data were acquired in mobility mode with a variable wave velocity using a linear ramp that started at 2500 m/s and ended at 400 m/s, a wave height of 40 V and IMS gas flow of 90 mL/min. A wideband enhancement file was generated with DriftScope v2.8 (Waters Corp.) for each of the three charge states shown in Figure 5.2b. These wideband enhancement files were then used to synchronise the pusher to the mobility time of the fragment ions in order to boost the intensity of each charge state family within the final reconstituted MS/MS spectra.

5.2.2 Merging of Enhanced High Duty Cycle Data

As each wideband enhancement file represents a different fragment ion charge state family, a separate experiment using each calibration file must be performed. In order to generate a complete spectra containing the high duty cycle data for each charge state family it is necessary to recombine the fragment ion data to a single spectra. To examine the repeatability of precursor ion characteristics within technical replicates an analysis of a previously collected complex data set was conducted. The data consisted of a *Drosophila melanogaster* cerebrum tryptic digest and was composed of >8000 peptides recorded in three separate LC-MS/MS experiments. These experiments were run on the same day in sequential order. The precursor retention times and m/z values for identical peptides were inspected for variation between experimental runs. The result of this analysis is shown in Figure 5.3. It can be seen that for each precursor these values are highly reproducible between experiments. This enable the generation of tolerance parameters over which to recombine MS/MS peaks lists.

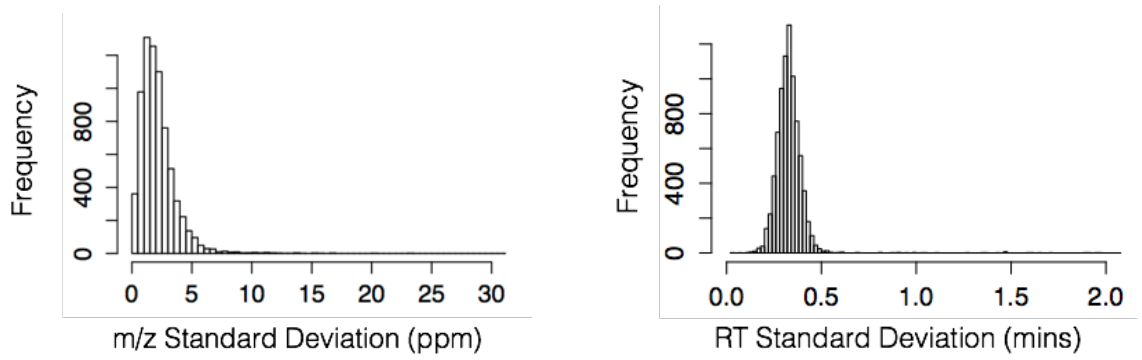


Figure 5.3: Evaluation of reproducibility for the analysis of technical replicates of a tryptic digest of *D. melanogaster* cerebrum. For each precursor ion that has been identified in all three of the triplicate runs the following information has been plotted. A) Histogram to show the standard deviation of each identified fragment ion measurement error in ppm. Measurement error across the runs for all precursor ions is below 8 ppm. B) Histogram to show the standard deviation of retention time (RT) measurement for each precursor. RT deviation is below 0.6 mins.

As a result of this analysis MGFMerge.py was written in Python to recombine fragment ion data from HD-DDA analysis of each nested charge state. MGFMerge takes three MGF files as inputs, one from each charge state calibration, and returns an MGF format file of the complete spectra for further analysis. In order for fragment ion data from the different files to be recombined into a single spectra, precursor ions must meet the following criteria: m/z within ± 7.5 ppm, retention time within $\pm$ 15 seconds and identical charge. In cases where the m/z of fragment ions is equal the highest intensity value is recorded.

The final HD-DDA experimental workflow is shown in Figure 5.4. Each fraction is analysed in three separate experiments using the +1, +2 and +3 wideband enhancement files, respectively. The samples were analysed using the LC-MS/MS parameters specified in Section 2.3 but with the addition of the mobility settings described above. The files were then processed in Protein Lynx Global Server v3.0.2 as previously described in Section 2.4. MGF files were then exported and merged using the *in house* Python script MergeMGF.py. The final merged MGF files for each fraction are converted to mzXML format and analysed by xQuest according to the workflow previously described in Section 3.3.2.

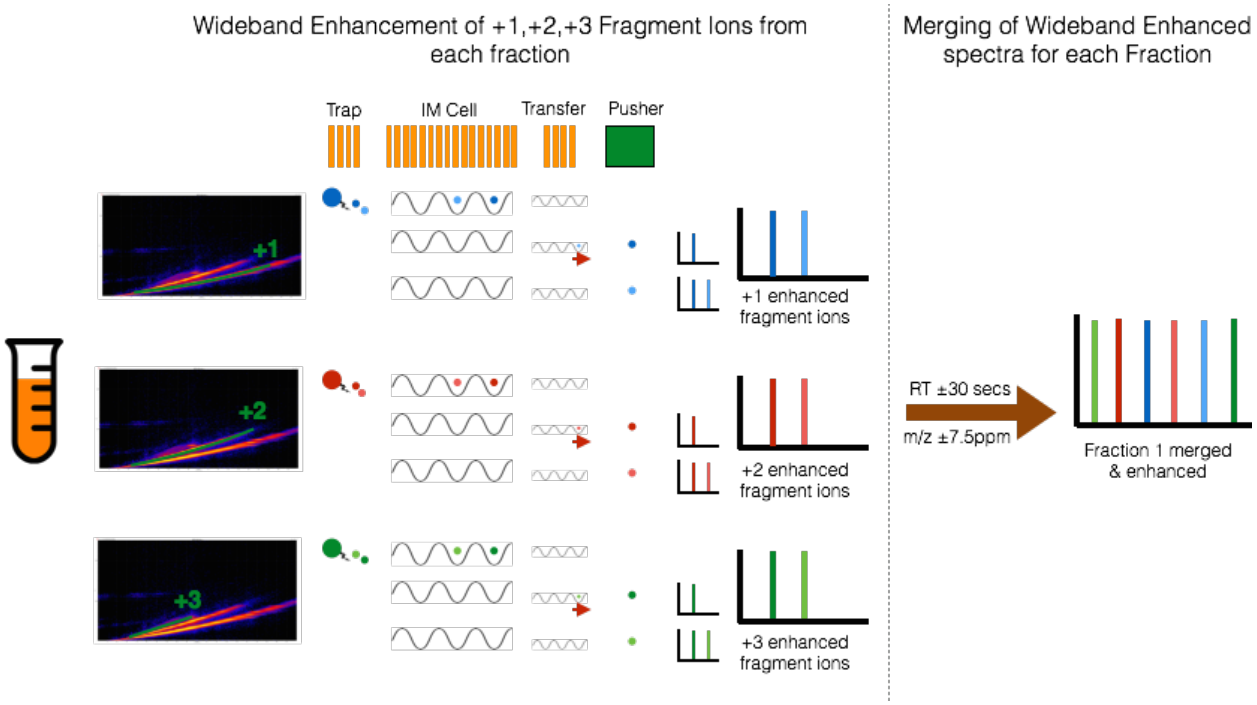


Figure 5.4: HD-DDA Experimental Overview. Following the creation of charge state calibrant files to synchronise the pusher pulse (See Figure 5.1B and 5.2B) each fraction is analysed using each of the charge state calibration files. Fragmentation of the precursor ions occurs in the Trap using the optimised collision energy ramp described in Chapter 3. Fragment ions are separated in the IMS cell using a variable wave velocity linearly ramped from 2500 m/s to 400 m/s over the course of a scan. IMS separation is maintained in the Transfer. The pusher pulse is synchronised by the calibrant file such that only fragment ions of a particular charge state enter the ToF for analysis. After MS/MS analysis the MGF files from each charge state calibrant file are merged to create one file based on the criteria described in Section 5.2.2. This file represents the final raw data file containing the enhanced duty cycle experiment for each charge state. This file is then analysed in xQuest according to the method described in Chapter 3.

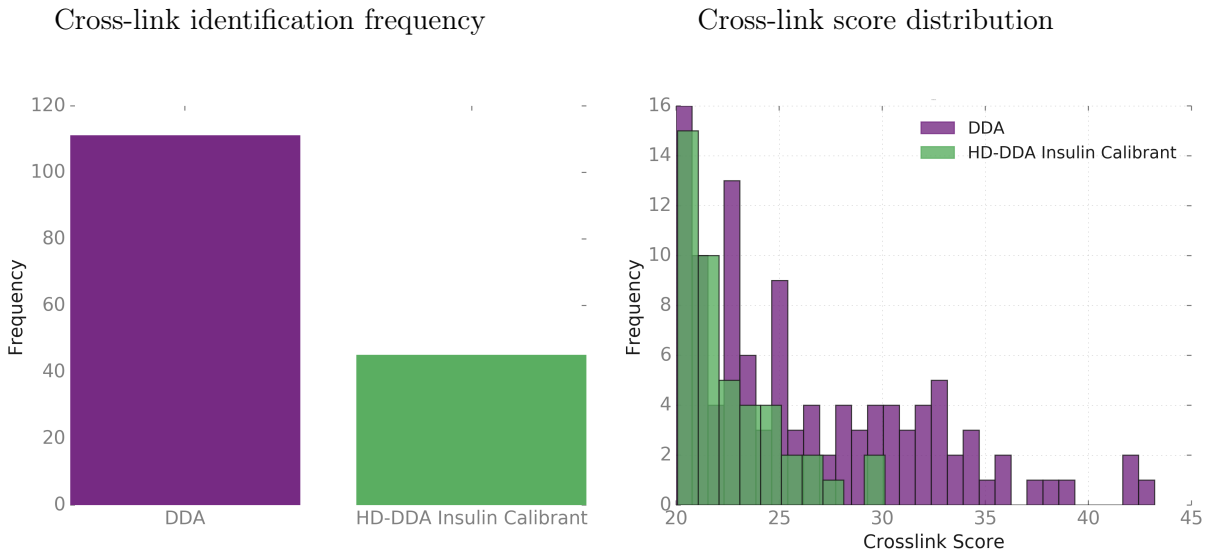
5.3 Results and Discussion

5.3.1 HD-DDA Analysis of Cross-linked BSA with Proinsulin C-Peptide Wideband Enhancement

In order to increase the duty cycle of the QToF mass spectrometer fractions of cross-linked BSA were analysed by the existing DDA methodology (as described in Section 3.3.2) and the HD-DDA method described in Section 5.2. Although the HD-DDA method requires a

number of repeat injections of the same sample, one for each wideband enhancement file, the synchronisation of the pusher pulse means that all ions exiting the mobility cell at times outside of the calibration do not enter the ToF for analysis. Therefore, in order to fairly compare cross-link identifications a single DDA experiment was performed.

A comparison of the number of the cross-links identified by xQuest can be seen in Figure 5.5a. The cross-links reported here score above 20 and have been validated in the raw data for the presence of a precursor doublet equal to the mass shift due to the cross-linker (Section 3.3.1).



(a) Comparison of BSA cross-link frequency between DDA (purple) and HD-DDA (green).

(b) Comparison of the BSA cross-link score distribution for DDA (purple) and HD-DDA (green).

Figure 5.5: Comparison of the cross-link histogram and score distribution for DDA (purple) and HD-DDA (green). Number of unique cross-links identified by sequence, including modifications such as oxidised methionine residues, that have been validated according to Section 3.3.1. A) Histogram of identified validated cross-links for DDA and HD-DDA method using the proinsulin C-peptide calibrant as shown in Figure 5.4. B) xQuest score distributions for the identified validated cross-links for the DDA and HD-DDA method using proinsulin C-peptide calibrant. The HD-DDA method provides fewer cross-link identifications with lower xQuest scores.

The number of cross-links identified by the xQuest analysis of HD-DDA data are greatly reduced from those observed in the equivalent DDA experiment. Only 46 cross-links were reported by the HD-DDA method compared to 131 by the DDA method. As the HD-DDA

method is expected to increase the duty cycle to improve fragment ion intensity a lower cross-link identification rate is unexpected. Due to the nature of the xQuest scoring algorithms (described in detail in the Introduction) higher intensity fragment ions should improve the overall cross-link score. To explore this further a histogram of the cross-link scores was generated for both datasets (Figure 5.5b). This comparison reveals that the quantity of cross-links with an xQuest linear discriminant score of twenty or above are comparable however, the HD-DDA method does not identify any cross-links that score above thirty.

The charge carried by cross-linked precursors is greater than linear peptides. Indeed 97 of 131 cross-links identified by the DDA method are charge state +4, with only 16 and 18 having charge states of +3 and +5 respectively. For the majority of these cross-link identifications, the fragment ions generated after collision induced dissociation(CID) will most likely carry a charge of +3 and below. For these ions the proinsulin C-peptide is an adequate calibrant for wideband enhancement. Although a calibration file for the +3 charge state family is generated the signal intensity for the region is very low. For 14% of the identified cross-links the timing of the pusher pulse may not match the exit time of these ions. Without the full spectra of fragment ions xQuest will not perform as favourably on the data.

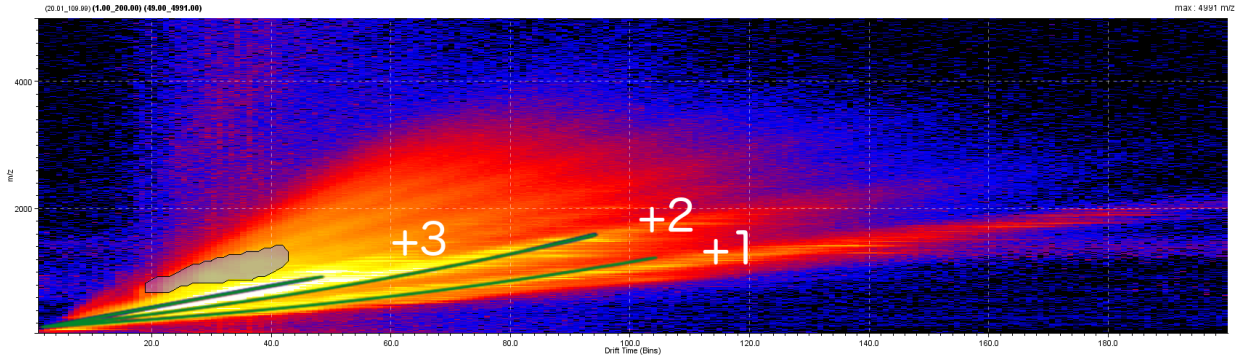


Figure 5.6: Mobility pattern for cross-linked BSA fragment ions analysed without wide-band enhancement (i.e. using the HD-DDA method without the use of a calibrant file to synchronise the pusher thereby allowing all ions to pass through to the ToF for analysis.). The proinsulin C-peptide calibration files have been superimposed over the mobility pattern (green) and charge states for the calibration files are labelled (white). Region containing +4 cross-linked ions indicated in light blue. The proinsulin C-peptide does not represent the full range of fragment ion charge states present in the BSA cross-linking experiment.

To explore the effects of the proinsulin C-peptide calibration a comparison of the mobility of both the insulin peptide and the sample was carried out using DriftScope v2.8 (Waters

Corp). The same cross-linked sample was analysed with the HD-DDA method but without pusher synchronisation. The wideband enhancement rule file was then opened within DriftScope and overlaid onto the acquired data. In this way the mobility pattern of the sample fragment ions was compared with the wideband enhancement file generated from the proinsulin C-peptide (Figure 5.6).

The calibration lines drawn from the nested charge states of the proinsulin C-peptide (green) lie over the cross-linked BSA mobility plot for the +1, +2 and +3 fragment ions. However, the existence of charge state families above +3 is also observed (Figure 5.6). Upon further investigation using DriftScope v2.8 an area within these charge state families was found to contain solely cross-linked fragment ions carrying a +4 charge (indicated in light blue on Figure 5.6). As this type of ion still contains the cross-linker a doublet distribution was visible and is shown in Figure 5.7. This region lies outside of the calibration lines created from the proinsulin C-peptide and was consistently observed between fractions. In order to include this region a more adequate calibrant is needed to synchronise the pusher. Further exploration of the nested charge state families outside of this area yielded no signal from either cross-links or peptides.

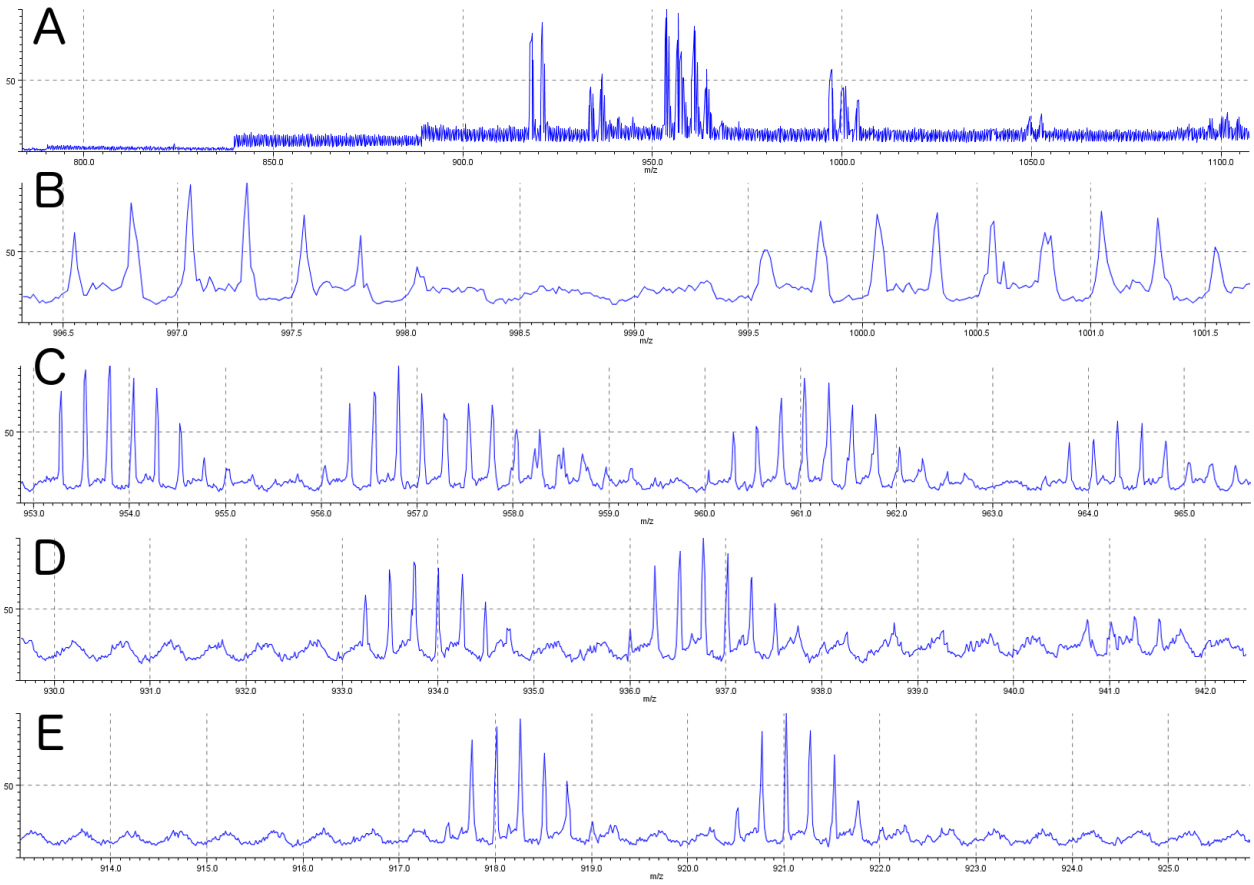


Figure 5.7: Analysis of highlighted region (light blue) in Figure 5.6 from cross-linked BSA fragment ions. A) All peaks found in region. XL ions at B) 996.6 m/z and 999.6 m/z C) 953.2 m/z and 956.2 m/z, 960.2 m/z and 963.8 m/z D) 933.2 m/z and 936.2 m/z E) 917.6 m/z light and 921.6 m/z. Four cross-linked fragment ions with a charge state of +4 have been identified in this isolated region. As the proinsulin C-peptide does not include +4 fragment ions these ions have not been directed into the ToF for analysis and will not have been included in the final MGF file that is searched by xQuest.

5.3.2 HD-DDA Analysis of Cross-linked BSA with Sample Wideband Enhancement

In order to generate a more accurate synchronisation of the pusher pulse for all fragment ion charge states, a wideband enhancement file was generated from the charge states families within the sample itself. As no signal was distinguishable from the nested charge states above +4 the final calibration files comprised +1, +2, +3 and +4 (Figure 5.8).

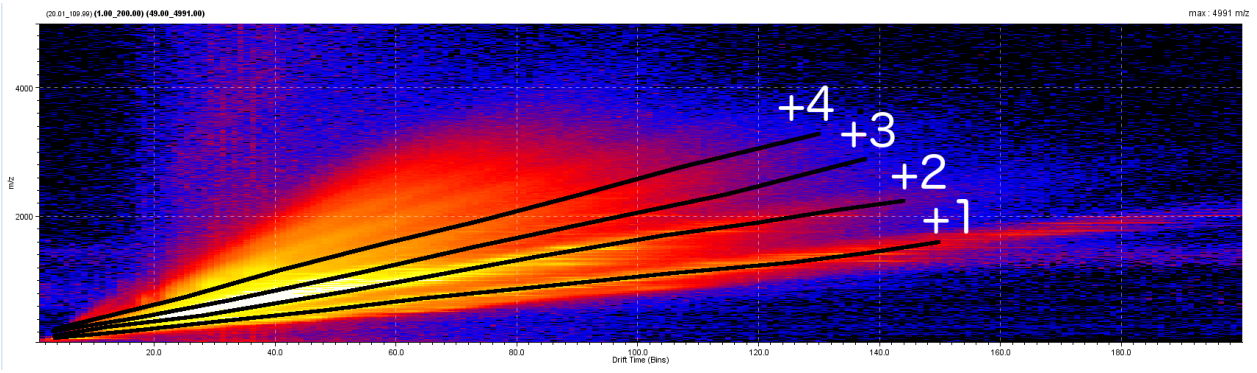
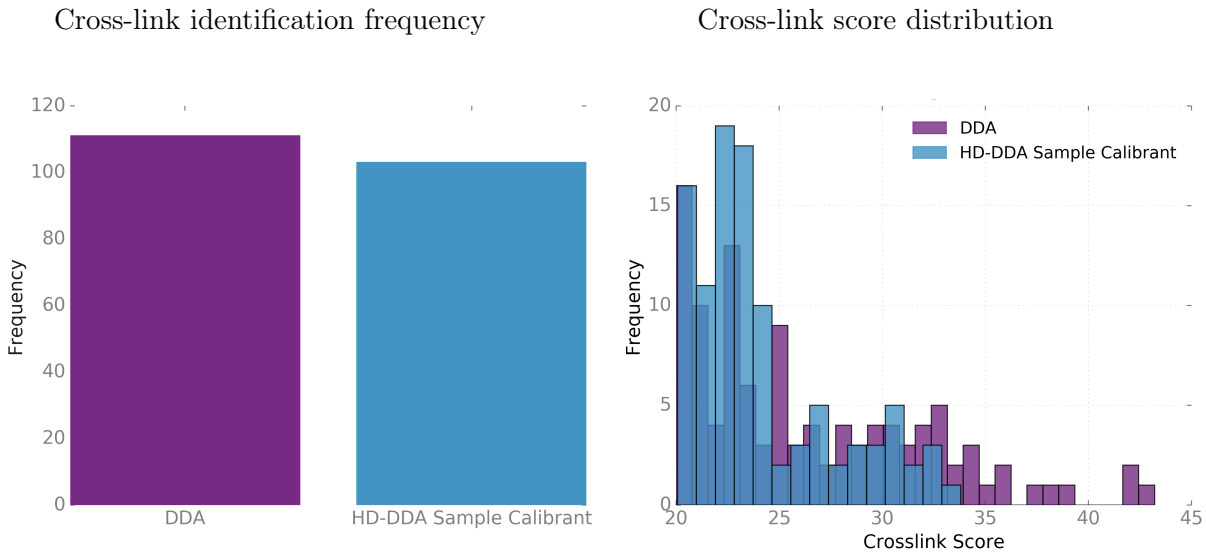


Figure 5.8: Wideband enhancement file generated from the mobility pattern of the cross-linked BSA sample. The BSA sample was analysed using the HD-DDA method without wideband enhancement. Calibration files for the +1 to +4 charge state fragment ions have been generated. Calibration lines are shown in black, charge states are labelled in white. Image generated using DriftScope v2.8. Intensity threshold values Min=30% and Max=100% counts using a logarithmic map intensity scale.

The cross-linked BSA sample was analysed four times using each of the generated calibration files to synchronise the pusher pulse across all fragment ion charge state families. Figure 5.9a and 5.9b show that this adaptation has improved the number of identified cross-links and the distribution of xQuest linear discriminant scores. With the addition of the sample calibrant 103 validated cross-links were identified with twelve of these scoring above 30. In comparison to the standard DDA method however, HD-DDA does not provide an overall increase to cross-link identification rates or scores.



(a) Comparison of cross-link frequency between DDA (purple) and HD-DDA with sample adaptation (blue).

(b) Comparison of the xQuest linear discriminant score distribution for DDA (purple) and HD-DDA with sample adaptation (blue).

Figure 5.9: Comparison of the cross-link histogram and score distribution for DDA (purple) and HD-DDA with sample calibrant files (blue). Number of unique cross-links identified by sequence, including modifications such as oxidised methionine residues, that have been validated according to Section 3.3.1. A) Histogram of identified validated cross-links for DDA and HD-DDA method using the BSA sample calibrant as shown in Figure 5.4. B) xQuest score distributions for the identified validated cross-links for the DDA and HD-DDA method using BSA sample calibrant. The BSA calibrant files have improved the number of cross-link identifications and scores, however the DDA method still provides a greater number of identifications with higher xQuest scores.

The wideband enhancement file requires the assignment of each drift time bin to a particular m/z value reported to four decimal places. Examination of the mobility plot for the cross-link sample (Figure 5.6) shows that the areas of greatest ion intensity covers a wider m/z range than in both the GFP and proinsulin C-peptide mobility patterns (Figure 5.1b and 5.2b, respectively). This broadening is particularly observed for the +2 and +3 charge state families. This suggests that although cross-linked fragment ions overlap with linear ions subtle differences in mobility may exist. It may therefore be more suitable to calibrate the pusher over a m/z range for each drift time bin.

5.3.3 Role of HD-DDA in Duty Cycle for both Calibrants

To ascertain the extent to which the wideband enhancement was successful in improving the duty cycle the spectral acquisition rate from the DDA experiment was compared to the rate observed in both the HD-DDA methods. In the original method development work by Helm et al. [44] the HD-DDA method showed a 60% higher spectral acquisition rate than the standard DDA.

In Figure 5.10 the mean number of MS/MS spectra recorded per minute across five minute retention time windows is shown. For clarity the results for the HD-DDA insulin and sample methods have been normalised to the acquisition rate for the DDA analysis. The sample calibrant shows a significant improvement in acquisition rate over the insulin calibrant. The duty cycle of the instrument has therefore been improved with the addition of wideband enhancement using the sample as a calibrant. This improvement however, does not lead to better cross-link identification rates.

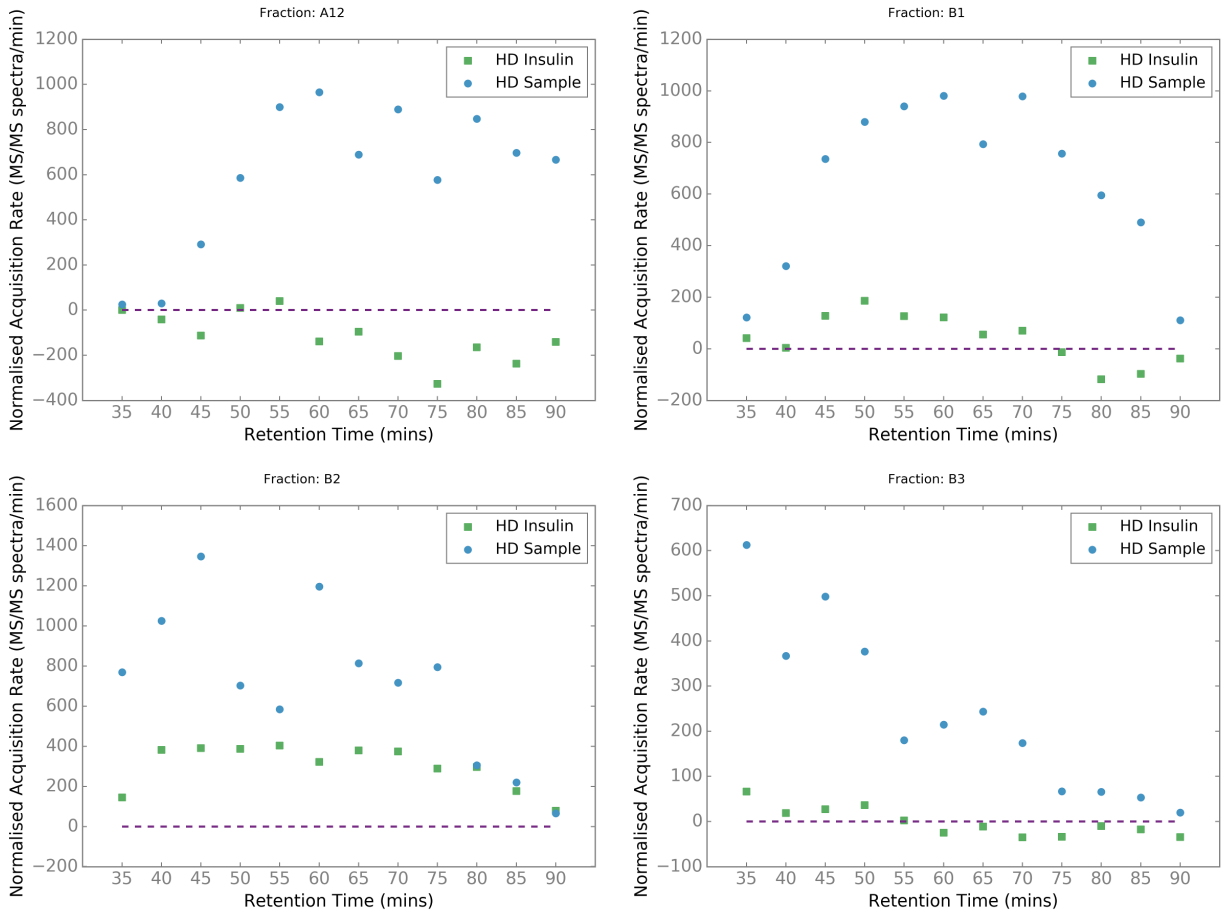


Figure 5.10: Assessment of the effect of each calibrant on the duty cycle of the instrument. The MS/MS spectra acquisition rate for the final combined charge state enhanced duty cycle HD-DDA experiment for both the proinsulin C-peptide and the BSA sample calibrant have been compared. Data have been normalised to DDA values HD-DDA with proinsulin - green square, HD-DDA with sample - blue circle and DDA - purple dashed line. Number of MS/MS spectra acquired for each 5 min retention time bin are shown. The BSA sample calibrant shows an improved acquisition rate compared to both the DDA and proinsulin C-peptide HD-DDA methods.

5.3.4 Comparison of Spectral Quality Across all Methods

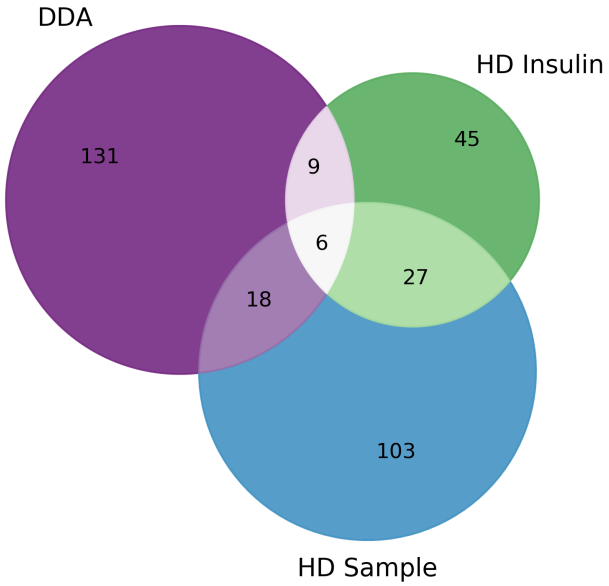


Figure 5.11: Cross-link residue pair overlap for DDA and both HD-DDA methods. Cross-links have been counted by unique residue position in the protein sequence and do not include those with sequence modifications such as oxidised methionine residues. Minimal overlap can be seen across each of the experimental methods.

In order to evaluate the effect of the HD-DDA method on individual cross-link fragment ion spectra the subset of unique cross-link residue pairs present in all methods was generated (Figure 5.11). The overlap in cross-link identifications was very poor with only six cross-links identified in all three. To identify possible trends between the effects of each method on the xQuest scoring algorithms the subscores for the overlapping set of cross-links were plotted (Figure 5.12).

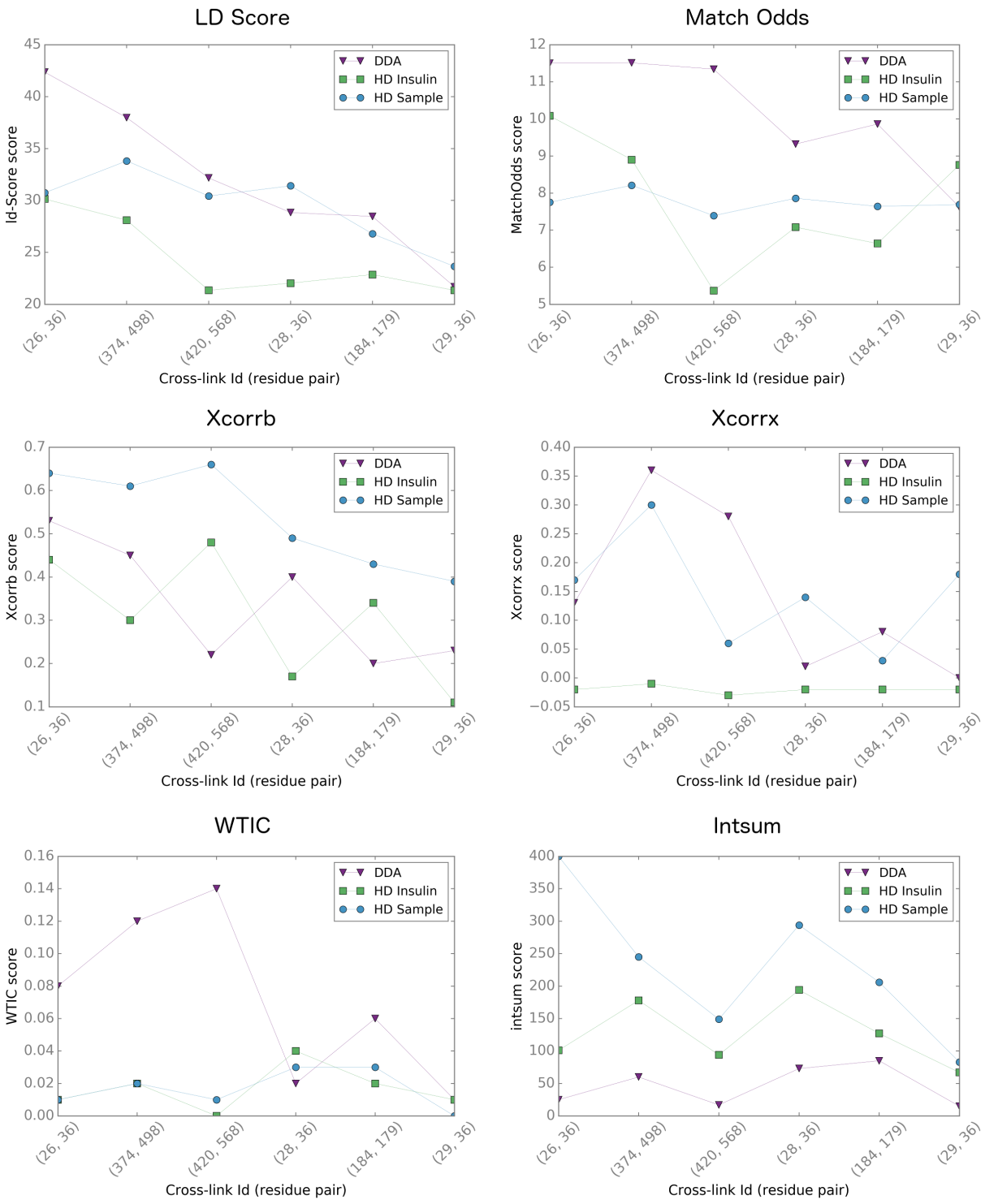


Figure 5.12: Comparison of the xQuest subscores for the cross-links that were identified in all methods. Cross-link residue pair has been shown and cross-links are plotted in order of total residue length. Subscores for DDA method shown as purple triangles, HD-DDA with proinsulin C-Peptide shown as green squares and HD-DDA with BSA sample calibrant are shown as blue circles. The score representing the sum of the spectral intensity and the linear fragment ion correlation are improved for the HD-DDA method with the BSA calibrant.

For all cross-links in this set the MatchOdds, WTIC and XcorrX subscore values show consistent improvement when using the DDA method. Linear fragment ions and spectral intensity however, produce the opposite effect. All cross-links in the set posses higher Intsum scores for both HD-DDA methods compared to the DDA. As this score is defined as the sum of the intensity of all peaks in the MS/MS spectra, the increase in fragment ion intensity is likely due to an improved duty cycle. This increase however, does not lead to an increase in cross-link identifications.

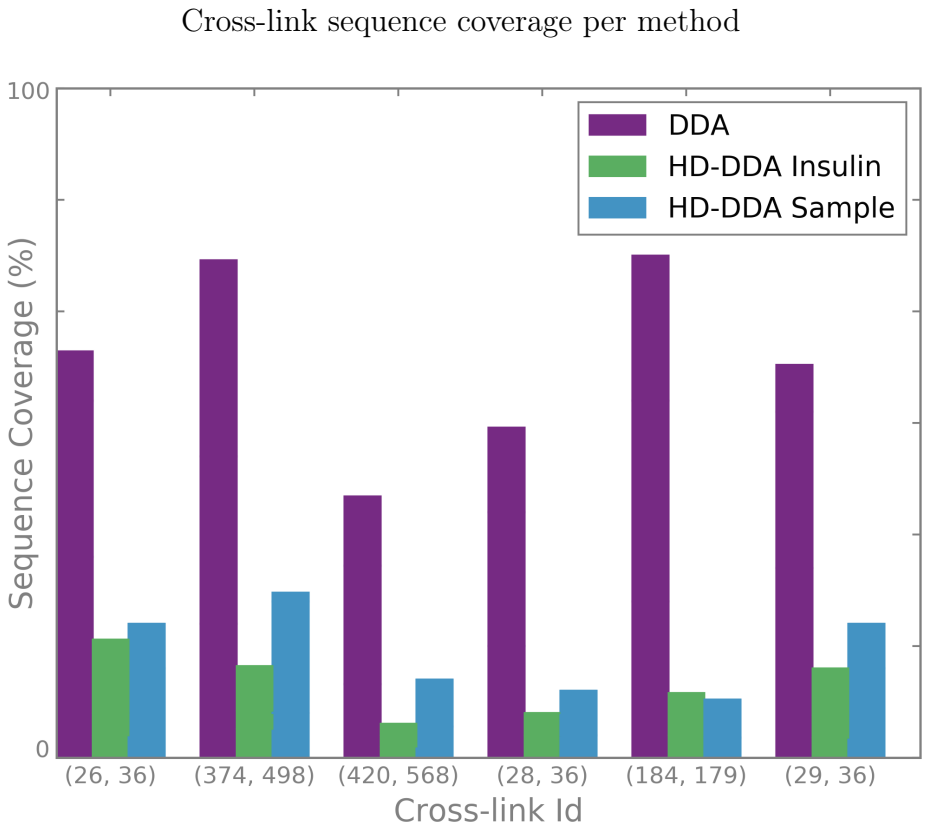


Figure 5.13: Percentage sequence coverage for each of the cross-links identified in all methods. Sequence coverage calculated based on the percentage of annotated ions from the theoretical maximum. Theoretical ion calculation is described in to Appendix E. Percentage sequence coverage for DDA method shown in purple, HD-DDA with proinsulin C-peptide calibrant shown in green, HD-DDA with BSA sample calibrant shown in blue.

The final xQuest linear discriminant score for each cross-link is given by a weighted sum of all the subscores. In almost all cases the DDA method generates higher final scores than either of the HD-DDA methods. MatchOdds has a mean contribution of 52% to the linear discriminant score.[114] As the largest contributor to the final score it is mostly responsible

for the improved performance of the DDA. It is also one of only two subscores which uses annotated peaks. All other xQuest subscores are based on the set of peaks which have been identified in both the light and heavy spectra, irrespective of an annotation. Figure 5.13 shows that the number of annotated peaks in the spectra for each cross-link is significantly greater for the DDA method than either of the HD-DDA methods. This suggests that although the HD-DDA methods may improve spectral acquisition rate, xQuest is not able to annotate the spectra as readily.

For the DDA and HD-DDA methods fragmentation is conducted in the trap using identical collision energy ramps. Hence differences in fragmentation patterns should be negligible. In addition, as stated in Section 3.3.4, xQuest does not account for all possible fragment ion classes. This should not however, affect the present analysis as all results have been subject to an identical set of search parameters by the xQuest algorithms. Consequently the most likely explanations for the reduction in annotations include: poorly synchronised pusher pulse or incorrect recombination of the data during the merging operation. In both cases a series of analyses were conducted prior to implementation to ensure these parameters were set for optimal performance. It is however, possible that fluctuations in instrument performance may require a more dynamic approach to parameter adjustment.

5.4 Conclusion and Further Work

In this work we have adapted the wideband enhancement method originally employed to dynamically synchronise the pusher pulse to the arrival of fragment ions at the entrance to the oa-ToF. To adapt this method to cross-link analysis multiple calibration files are necessary to accommodate the arrival times of ions with a specific m/z across different charge state families. In doing so the peaks from the spectra are split across multiple experimental files. We have constructed an algorithm that recombines these peaks based on precursor ion characteristics that persist over multiple analyses.

We have shown that using a linear peptide calibrant to synchronise the pusher in the ToF provides an insufficient model for the behaviour of cross-linked digest samples in mobility space (Figure 5.6 and 5.7). To overcome this we present a further adaptation which uses the

sample itself to synchronise the pusher pulse, incorporating higher charge state families of fragment ions (Figure 5.8).

Although there is some evidence to suggest that the duty cycle has been increased by the addition of wideband enhancement (Figure 5.10) a corresponding increase in cross-link identification rates has not been observed (Figures 5.5a and 5.9a). There is limited overlap between cross-link identifications by the DDA and both wideband enhancement methods. For those cross-links that have been identified in all three experiments a significant reduction in the annotation of fragment ions is observed for the HD-DDA methods (Figure 5.13.

Analysis of the mobility plot for the fragment ions in a cross-linked BSA sample suggest that a wider m/z range may be necessary to account for both the cross-linked and linear fragment ions. This motivates the usage of a calibration method such as that describe in Section 4.3.3 which may provide better results. It should however be noted that this method may cause a reduction in the improvement to the duty cycle as it uses a m/z range to synchronise the pusher pulse. This range may contain noise in addition to the ions of interest.

The most likely explanation for the decrease in peak annotation is the recombination of fragment ion peaks by MGFMerge.py. The data used to identify the tolerance parameters represent only linear peptides. In addition, temperature fluctuations and solvent concentrations impact the retention time profile of eluting peptides. Furthermore, the calibration files were generated using the mobility plot for the fraction containing the highest number of cross-link identifications. As each fraction represents a different distribution of peptide sizes it may be necessary to generate a calibration file for each charge state family per fraction. The method however, already requires the generation of 4 calibration files and that the experiment be run four times for each fraction. Consequently it necessitates the use of higher sample volumes and additional analysis time. Although increasing the variation of calibrant files does not increase the time required to analyse the data it will lead to increased method complexity. Once biological and technical repeats are taken into consideration the increased duration of HD-DDA analysis of cross-links would require a significant increase in identification rates to justify the utility.

Chapter 6

Computational Solutions for the Analysis of Cross-linked Peptides

6.1 Introduction

6.1.1 Evolution of Crosslinking as a Structural Technique

Over the last twenty years cross-linking mass spectrometry has become an ubiquitous presence in the literature (Figure 6.1). Presently the Web of Science analytics portal reports 2,941 publications.[85] As discussed in Section 1.6.1 an ever expanding range of chemical cross-linkers has enabled the covalent modification of many different amino acid chemistries.[100,102,101,117,54,43,74] Improvements in mass spectrometer design along with analyser sensitivity and resolution have played an important role in improving the accuracy of identifications.[76,25,38,35] In addition, the combination of Solid Phase Extraction (SPE) and Strong Cation Exchange (SCX) into a single fractionation step improved the enrichment process by reducing the number of purification steps required. This resulted in fewer fractions for MS analysis and reduced sample loss.[92]

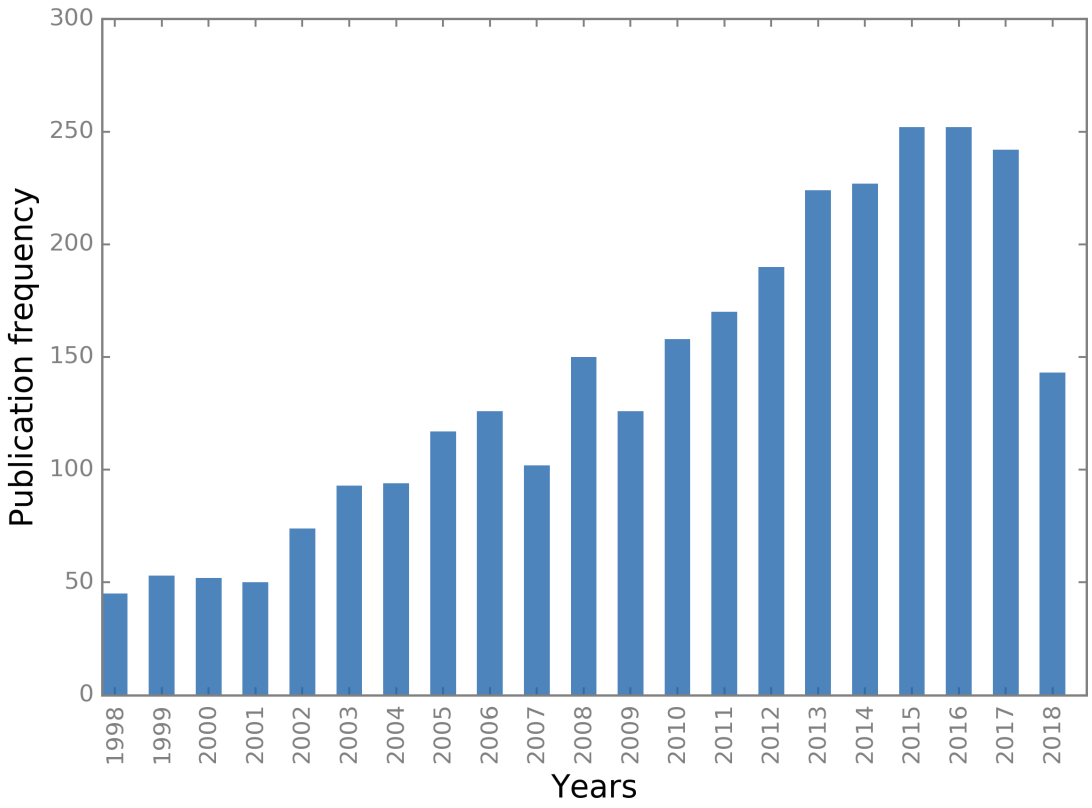


Figure 6.1: Total publications containing cross-linking mass spectrometry as a topic over the last twenty years. Graph produced using Web of Science.

The development of cross-linking protocols[59] and the advent of user friendly software applications[87,63,40,116] has enhanced the accessibility of the technique. Since 1998 ninety five publications relating to cross-linking mass spectrometry software can be found in the literature. As discussed in Section 1.6.3 many of these have been deprecated, withdrawn or are no longer supported. For those that remain varying levels of support exist for the user. Most require some level of computational expertise to install and execute.

Despite the extent of innovation in the field, there have been few guidelines offered for the analysis and validation of cross-linking identifications. To ensure cross-links are genuine substantial care must be taken in the analysis of fragment ion spectra. As discussed in Chapter 1 statistical scoring methods are not always reliable. Arbitrary scoring thresholds were are necessary but not sufficient to distinguish a true identification. In addition, these scores are often not reported in correspondence. Cross-link validation therefore, is often subjective and vulnerable to bias. The most common method of cross-link validation is to evaluate the

lengths of the identifications on a three dimensional structure.[61,16,99,29] As previously stated this often involves calculating the Euclidean distance between the carbon α residues of each amino acid in the cross-link. This shortest path often penetrates the surface of a protein and is therefore not a true representation of cross-link length. Furthermore this method of validation requires an atomic resolution structure of the protein under investigation. Hence this method serves only to add confidence to cross-link identifications rather than to confirm them.

In order to properly evaluate cross-link identifications it is necessary to interrogate MS/MS spectra. For this analysis to combine robustness and accessibility to the wider structural biology community, a protocol describing this process must be developed. The cross-linking community has begun to make some advancements in this area. Iacobucci and Sinz [47] published a set of guidelines to avoid mis-assignments of cross-links in MS data. A recommendation for the use of high mass accuracy was made for both the MS analysis and the subsequent database searches. To avoid inaccurate assignments of cross-link:spectrum matches a peak matching maximum tolerance of 5 ppm for precursor ions and 10 ppm for fragment ions was suggested. In addition to instrument and data parameters, two further guidelines for the validation of spectra were presented: consideration of Signal to Noise Ratio (SNR), defined as the ratio of assigned and observed peaks, and fragmentation of both the peptides in the cross-link.

High sequence coverage of both peptides is necessary to confirm the identity of each peptide and to unambiguously determine the position of the cross-linker. The xQuest/xProphet software application contains a score based on the intensity of ions from either peptide: the WTIC score. This score and the MatchOdds score are the only two subscores that consider solely the annotated peaks. In addition, xQuest has no score to describe the SNR of an identified cross-link:spectrum match. In further investigations it was found that the spectral image presented in the user interface for xQuest only contains a subset of the observed peaks, those that are found in both the heavy and light MS/MS spectra. If this subset does not include the base peak these spectra may also be renormalised to the most intense ion in the subset. In addition, xQuest does not consider all types of fragment ions. As discussed in Chapter 3, this includes immonium ions, diagnostic BS3 ions and cross-linked fragment ions

with fragmentation events on both peptides. Furthermore, the annotated ions are only reported in a transient Graphical User Interface (GUI), through a local web server accessed via a web browser. These are created in a temporary directory that is overwritten upon opening each cross-link displayed in the specific results page of this user interface. This prevents a high throughput batched analysis of the annotated peaks without significant computational scripting experience.

To address this two computational tools were developed to aid in the validation of cross-linked spectra. The first, ValidateXL, serves as a quality control addition to the xQuest/x-Prophet software. Manual validation is directed to where it is of most benefit on the basis of sequence coverage for each of the peptides in the cross-link. The second, AnnotateXL, offers command line functionality that annotates an observed list of m/z peaks and intensities based on calculated theoretical peaks for a given cross-link identification. It considers the aforementioned ion types that are not included during xQuest analysis and provides a measure of SNR for a candidate cross-link:spectrum match.

6.2 Materials and Methods

6.2.1 ValidateXL.py

ValidateXL is written in Python 3.5 and is available to download from `https://github.com/ThalassinosLab/ValidateXL`. The software is offered under a GNU license allowing users to download the full source, make alterations and re-distribute without guarantee or warranty. The program has been designed to extract additional information from the generated xQuest results files. Following cross-link analysis of an experimental data file by xQuest the cross-link results are presented to the user through the bespoke web interface. Cross-link identifications are displayed in a table format with the option to view the MS/MS spectra for all peaks that have been found in both the heavy and light scan. The theoretical ion series for the cross-linked and linear fragment ions is also generated with matched ions highlighted.

In addition to the information represented through the web interface, several other files containing pertinent information about each identification are also generated and stored in

the xQuest results folder of the search. The merged_xquest.xml file contains details of all the peptide spectrum matches that have been searched by the software. These results are then ranked according to the final linear discriminant score. Each match is an element that contains attributes describing the cross-link identification in greater detail than presented in the xQuest GUI.

In addition to sequence information about each match and the results of the different scoring algorithms the XML file also contains details of the number of annotated ions for both peptides. ValidateXL makes use of a standard Python library called ElementTree to extract details of the number of matched ions for each peptide. Matches corresponding to the highest ranking cross-link identification for a peptide spectrum match are converted to a table style format known as a dataframe using the open source Python library, Pandas. The sequence coverage for each peptide is then calculated and used to filter the cross-link identifications into three groups. The results are returned to the user in the form of three CSV files containing sequence and position information for each cross-link as well as the sequence coverage. A schematic representation of the algorithm is presented in Figure 6.2.

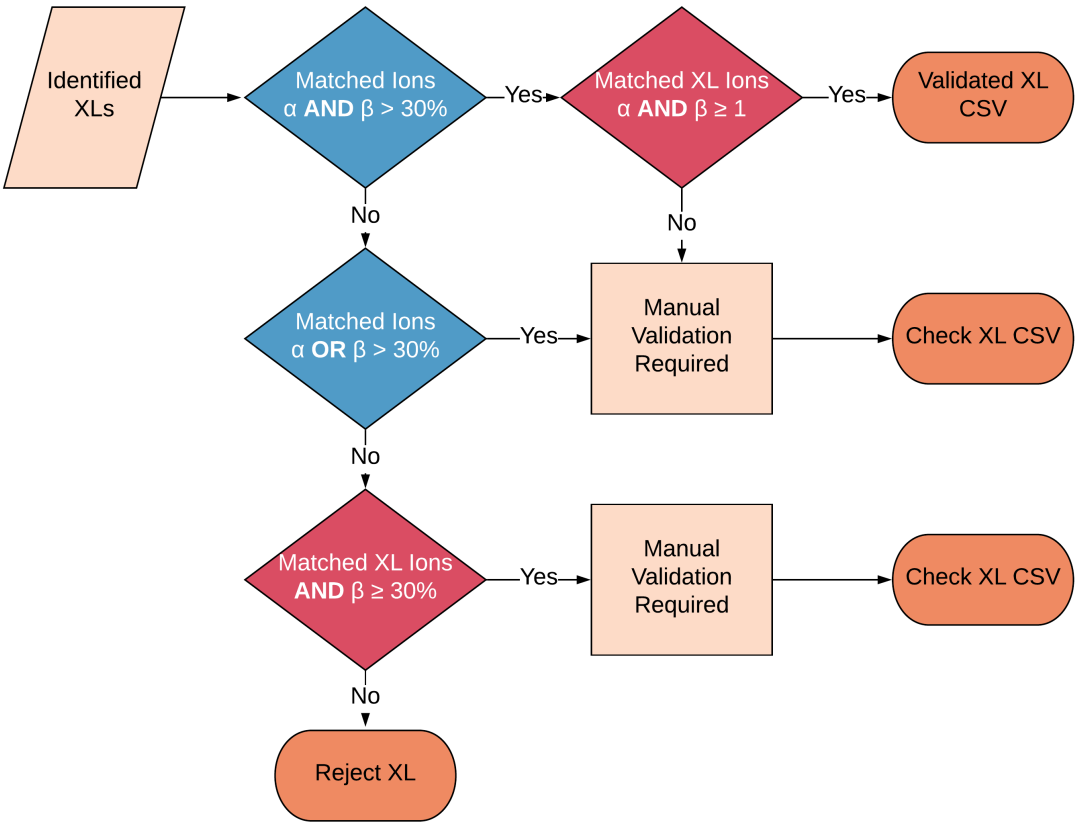


Figure 6.2: Schematic representation of the ValidateXL.py algorithm. The algorithm interrogates the XML result file provided in the xQuest results folder following analysis of cross-linked data. To determine sequence coverage the annotated linear and cross-linked fragment ions are considered separately. A full description of the calculation of sequence coverage is presented in Appendix E. Following execution three CSV files are returned; automatically validated cross-links, cross-links of an acceptable standard but in need of manual validation and cross-links which display such poor sequence coverage that they can be rejected.

ValidateXL classifies cross-link identifications based on sequence coverage. For a cross-link to be considered valid they must have at least one cross-linked fragment ion and 30% sequence coverage for linear fragment ions on both the α and β peptides. Rejected crosslinks have an overall sequence coverage of less than 30% across all fragment ions. The cross-links recommended for further manual validation are the remaining set: those which have more than 30% sequence coverage on the α or β or at least 30% coverage of the crosslinker ions. For the latter group manual inspection of the MS/MS spectra is highly recommended. Appendix E contains more information on the determination of the correct sequence coverage threshold.

6.2.2 AnnotateXL

Many different software applications exist for the analysis of cross-linked datasets.[116,111,63,39] As yet there is no tool to independently assess the quality of an assigned cross-link:spectrum match. AnnotateXL has been developed to provide this assessment. The software is written in Python 3.5 and is offered to the community under a GNU General Public License. It is available to download from `https://github.com/ThalassinosLab/Annotate\textunderscoreXL` and was written under an Object Oriented programming paradigm. A module diagram showing the connections between classes can be seen in Figure 6.3. AnnotateXL has been designed to work with deconvoluted MS/MS data.

The `Cross-link` base class extracts information about a cross-link from a string representation of the form: "αSequence-βSequence-an-bn" where n represents the position of the cross-linker. This class uses the `Peptide` base class, which yields a string representation of the amino acid sequence for all of the N and C terminal ion series. The `Peptide` class is itself dependent upon the `AminoAcid` base class to obtain the monoisotopic mass for a sequence from the `Utils.py`. `Utils.py` stores basic information in the form of hash maps. This includes monoisotopic masses for each amino acid and for the immonium ions and diagnostic ions.

The `Fragmenter` base class generates a list of theoretical fragment ions from a cross-link string input. The `CID` method on this class initiates the fragmentation process to yield all possible theoretical fragment ions. There are five types of fragment ion considered by the Fragmenter class when generating theoretical fragment ions for a cross-link: `XlFragmentIon`, `LinearFragmentIon`, `ImmoniumFragmentIon`, `PrecursorFragmentIon` and `DiagnosticIon`. These sub classes inherit from an abstract base class `FragmentIon`. This serves as a skeleton to define a series of methods that are used by each of the subclasses. The abstract methods defined in this super class describe features that are necessary in the subclasses listed. These include the mass of the fragment, and a Roepstorff representing the nomenclature as discussed in Section 1.5.1. It also includes a sequence representation and an ion name describing the fragment generated.

The generation of theoretical fragment ions is described in detail in Appendix E. Briefly, the `Fragmenter` class generates string representations for the linear fragment ions based on

amino acids in the α and β peptides. As the C terminal ion series are numbered from the C terminal a series of string manipulations is carried out to generate the correct directional sequences. The linear fragment ion strings are generated by iterating through the sequence strings of both peptides up to the position of the linker. These strings are used by the `LinearFragmentIon` subclass to calculate the terminal mass additions and subtractions necessary for an a,b,c or x,y,z ion type. The final mass Roepstorff nomenclature and sequence are returned by this subclass.

The cross-linked fragment ions are generated using the same set of string manipulations. Each peptide is considered individually to create the full set of sequence strings. The set of linear fragment ions is then subtracted from the full N and C terminal α and β peptide strings. In this way only the ions containing the linker remain. To generate the complete list of ions with only a single fragmentation event on one of the peptides the cartesian product of the full length peptide and the list of strings for the other peptide is created. An example of this is displayed in Appendix E. The final element of this generated list is removed as it represents the precursor ion. For cross-linked ions where fragmentation occurs on both peptides the cartesian product of the list of strings for both peptides is created. Once these strings have been obtained the `XlFragmentIon` subclass uses the strings to generate mass values, nomenclatures and sequences for all ion types. These include the mass of the cross-linker (without the mass of the leaving groups lost during conjugation -Figure 1.14). If the theoretical ion has been generated through a single fragmentation event the `XlFragmentIon` subclass also adds a mass modification to correspond to the N and C terminal ends of the full length peptide: two hydrogen atoms and an oxygen atom. At present AnnotateXL has been designed to work with the DSS/BS3 cross-linker. It can however, be modified by changing the mass of the linker in this subclass.

The simplest of the subclasses is the `DiagnosticFragmentIon` subclass. It simply returns the mass of the diagnostic ions from the Utils module. The `ImmoniumFragmentIon` subclass provides information in a similar way, it returns the masses of all possible immonium ions based on the sequence of both peptides in the cross-link. The nomenclature is simply the single letter amino acid code prefixed with an "IM_".

Once all of the theoretical ions and associated masses have been calculated they must

be matched to those observed during an experiment. Annotated_Ions.py offers the user command line access to AnnotateXL. To execute the program two inputs are required. The cross-link string and a CSV file containing two columns titled m/z and intensity. This script executes the `CID` method on the `Fragmenter` class instance to generate all possible theoretical ions. The `ObservedIon` base class converts each observed ion to an object with a m/z and an intensity. The `Annotator` base class then creates a hash map that contains all of the observed ions and details of any matches to theoretical ions. This is output to a CSV file and used to generate a PNG graphic of the annotated spectra. Matched ions are coloured according to the fragment ion subclass: cross-linked in red, linear in blue, immonium ions in purple and diagnostic ions in green. Each matched peak is annotated above the peak intensity the nomenclature is explained in Appendix E. Unmatched ions are displayed in grey. All peaks in the MS/MS are normalised to base peak intensity, which is represented as 100%.

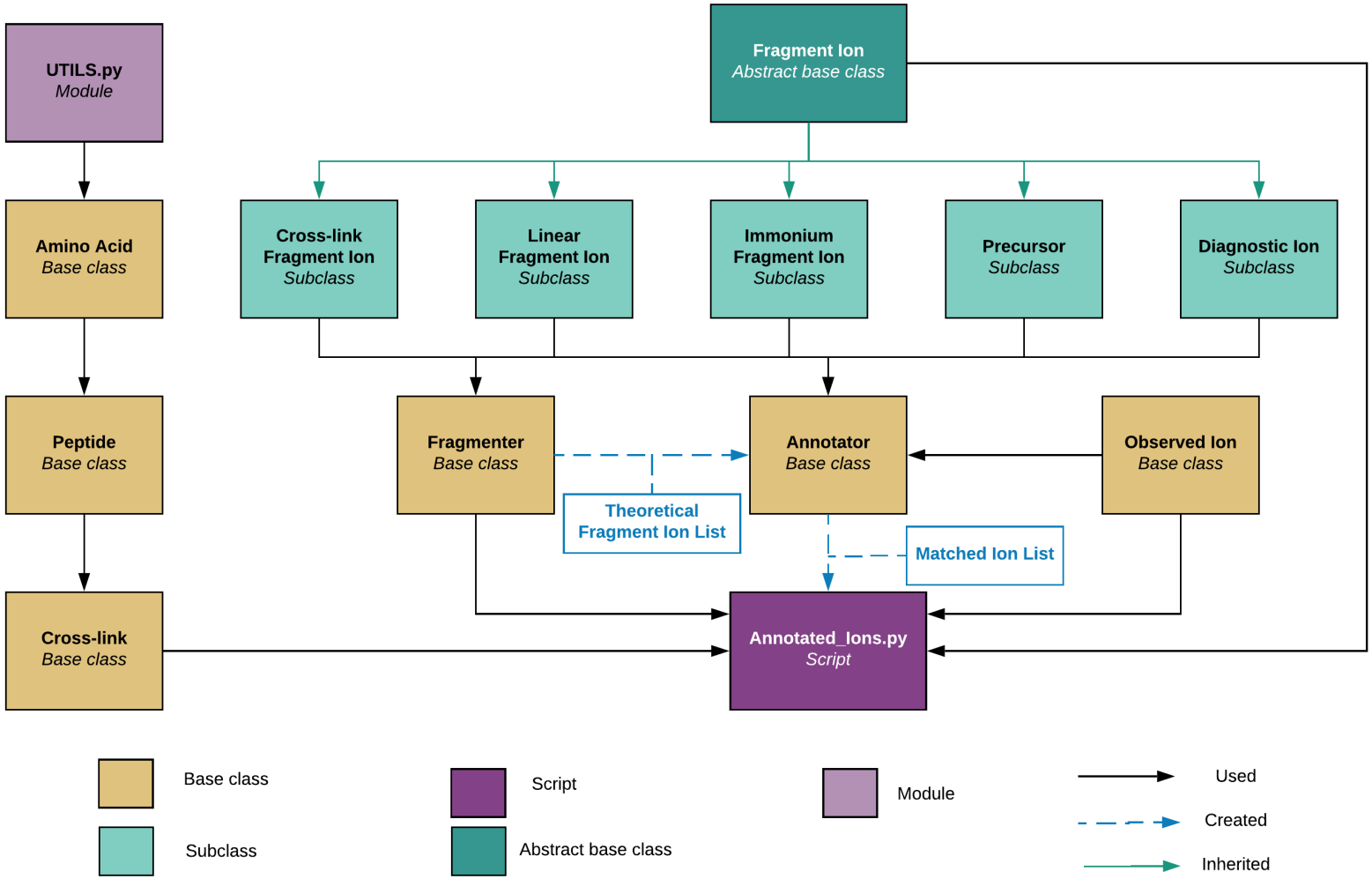


Figure 6.3: Schematic representation of the AnnotateXL application. AnnotateXL is an object oriented python programme described in detail above and in Appendix E. The diagram represents how each of the classes interact. The programme is executed by the Annotated_ions script (purple) and generates two lists: A theoretical fragment ion list and a matched ion list (blue). These allow calculation of signal:noise ratio and annotation of cross-linked peptide mass spectra. Arrows represent order of class execution rather than inheritance.

6.3 Results and Discussion

6.3.1 Analysis of DDA Datasets with ValidateXL

Effects of Validation by ValidateXL on Triplicate Datasets

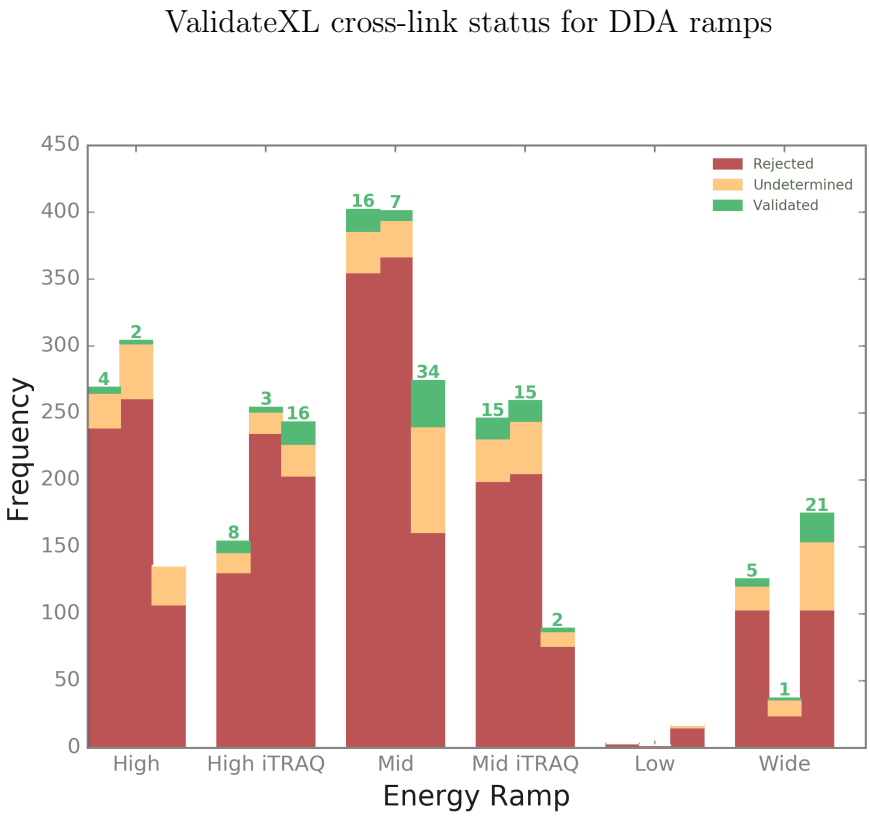


Figure 6.4: Cross-link status determined by ValidateXL for all unique BSA cross-links identified by xQuest. Rejected cross-links shown in red, those in need of manual validation in orange, automatically validated in green. The number of automatically validated cross-links is highest for the Mid energy ramp.

ValidateXL was set to run in a loop over all the experimental files in the triplicate DDA dataset. Execution took 80 seconds on a laptop with a 1.4 GHz Intel Core i5 processor and 4 GB of RAM. Figure 6.4 shows the breakdown of the results for all of the tested DDA ramps. Following automated validation the Mid ramp remains the best performing energy ramp with 16, 7 and 34 cross-links having a sequence coverage of at least 30% for both peptides and cross-linked ions. A disproportionate increase in the number of rejected cross-links was also observed (red bars). As xQuest does not account for sequence coverage this result was to be

expected. None of the spectra for cross-links in the rejected set presented sufficient SNR to be considered as genuine identifications.

Table 6.1 shows the final quantity of cross-link identifications from each experiment that were found to have suitable sequence coverage. The table separates these by algorithmic and manual determination. It also provides details of the number that were recommended for manual validation by ValidateXL. In comparison to the quantity reported by the xQuest search alone the number recommended for validation by ValidateXL is greatly reduced. For the best performing DDA ramp the quantity reduces from 103 to 31, 111 to 27 and 131 to 45 for each of the triplicate runs. This greatly decreased the amount of manual evaluation required. The biggest reduction was observed in the second HighiTRAQ analysis. In this instance the number of cross-links requiring validation was reduced from 155 to 16. During manual validation the most frequently observed reason for rejecting a cross-link:spectrum match was insufficient fragmentation of one of the peptides. Other reasons included insufficient fragmentation of the precursor or an absence of cross-linked fragment ions.

Table 6.1: Increase in validated cross-links following manual validation using ValidateXL.py. A large reduction in the total number of validated cross-links when compared to the number of cross-links scoring over 20 shows that a simple scoring threshold is not sufficient to determine a true cross-link identification. The reduction in the number of cross-links requiring validation reduces the time scale of a complete cross-linking experiment.

Ramp	Total validated cross-links	Algorithmically determined	Manually determined	Recommend for manual validation	Full identifications (Scoring >20)
High 1	10	4	6	26	270 (87)
High 2	16	2	14	42	301 (91)
High 3	7	0	7	29	136 (106)
HighiTRAQ 1	16	8	8	15	227 (69)
HighiTRAQ 2	18	18	0	16	250 (155)
HighiTRAQ 3	21	16	5	11	172 (81)
Mid 1	37	16	21	31	402 (103)
Mid 2	37	30	7	27	400 (111)
Mid 3	39	15	24	45	273 (131)
MidiTRAQ 1	29	15	14	33	259 (108)
MidiTRAQ 2	24	15	9	39	88 (41)
MidiTRAQ 3	23	18	5	24	246 (105)
Low 1	0	0	0	0	3 (1)
Low 2	0	0	0	0	2 (0)
Low 3	0	0	0	2	17 (9)
Wide 1	10	5	5	18	125 (44)
Wide 2	6	1	5	12	37 (23)
Wide 3	16	16	0	25	175 (99)

In order to assess the change in performance of the energy ramps following this validation, the mean and standard deviation of the cross-link identification rate was plotted for all of the DDA energy ramps (Figure 6.5). The Mid energy ramp once again remains the best performing. The MidiTRAQ ramp however, outperforms the HighiTRAQ once the full validated dataset is considered. As discussed in Chapter 1 the High energy ramp loses cross-linked peaks, most likely through the fragmentation of the cross-linker amide bond. For a cross-link to be categorised for further manual validation ValidateXL requires at least 30% sequence coverage for either the alpha, beta or cross-linked fragment ions. The mean XCorrx value for the intersection of cross-links validated for the HighiTRAQ and MidiTRAQ ramps was 0.251 and 0.316 respectively (Section 3.3.4 Table 3.4). Hence the loss of cross-linked peaks in the MS/MS spectra is the most likely reason for the increased performance of the MidiTRAQ ramp.

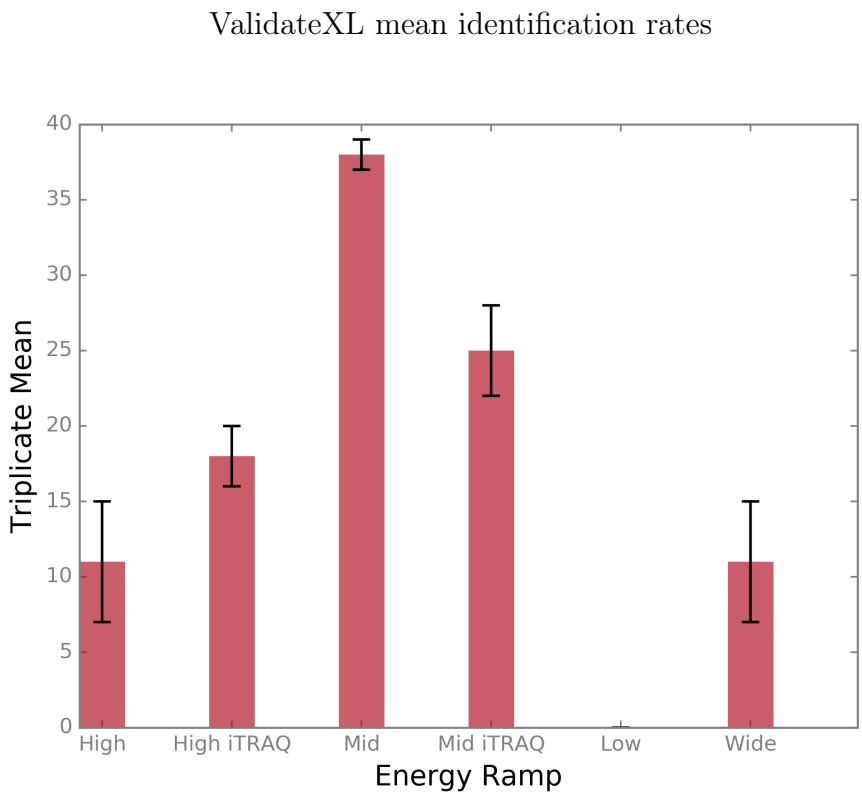


Figure 6.5: Mean for the unique BSA cross-links identified in the triplicate dataset following both automatic and manually validation in all DDA energy ramps. ValidateXL was used to validate the cross-links as described in Section 6.2.1. Error bars show standard deviation of the number of cross-links identified across the triplicate technical repeats.

ValidateXL mitigated extensive validation, in most cases reducing the number of cross-links requiring manual inspection by at least 50%. The final number of cross-links identified was substantially reduced when compared to a simple scoring threshold, from 103 to 37, 111 to 37 and 131 to 39 in the best performing DDA ramp. As the program also extracts information regarding sequence coverage it enable further more in-depth analysis to be carried out.

6.3.2 Effect of Validation and Energy Ramps on Fragmentation Efficiency for Alpha and Beta Peptides

The sequence coverage for each peptide in a cross-link was then analysed. Fragmentation of both the peptides in a cross-link has been observed to be unequal.[111,33,57,47] The larger peptide of the two frequently fragments more readily than the other resulting in a higher sequence coverage. Throughout this analysis we define the larger of the two peptides as the alpha peptide, in line with the xQuest definition. Figure 6.6 displays the mean percentage of annotated peaks for both the alpha (red) and beta (orange) peptides for the identified cross-links. Higher sequence coverage is indeed observed for the alpha peptide in most of the energy ramps. In published datasets, when using both CID[33] and HCD[57] with an Orbitrap analyser the beta peptide consistently displayed poorer fragmentation, with only 22% of the most intense annotated fragment ions corresponding to the beta peptide in CID analysis. Our analysis shows an improvement, with at least 40% sequence coverage for the beta peptide in all tested energy ramps.

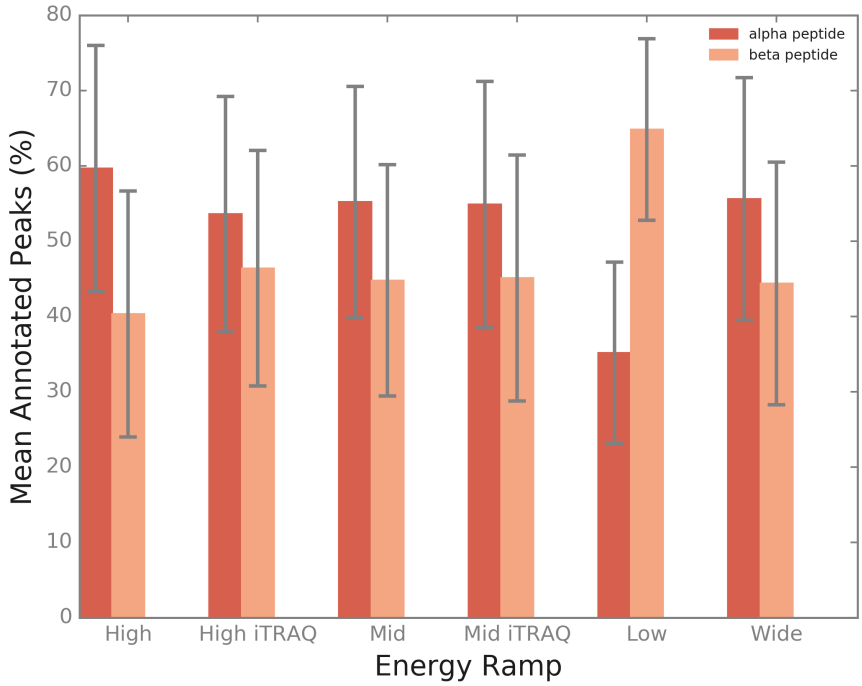


Figure 6.6: Mean percentage of annotated alpha and beta fragment ion peaks in MS/MS spectra of unique BSA intramolecular cross-linked peptides identified by xQuest. The height refers to the mean for each tested energy ramp. Error bars display the standard deviation. The sequence coverage was determined according to the method described in Appendix E. The beta peptide represents the shortest peptide by sequence.

The increase in sequence coverage may be explained by considering the calculation of collision energy within specific vendor instrument operation software. Normalised Collision Energy (NCE) applied by ThermoScientific compensates for the mass dependency on optimal collision energy by applying a linear percentage of the available energy for a particular m/z.[108] The Waters Corp. energy ramp exposes an ion to a range of energies over the course of the scan. This may be more advantageous for cross-links, since each peptide within the cross-link has a different m/z. The optimal fragmentation energy is therefore unlikely to be related solely to the precursor m/z but will differ for each of the peptides. Thus, a range of energies would likely be more optimal.

6.3.3 Effects of Validation by ValidateXL on QToF Experiments

In order to compare the performance of ValidateXL against the other methods of cross-link validation, ValidateXL and Jwalk analysis was used to evaluate all the QToF methods developed during this work (Table 6.2). The number of cross-links identified by Jwalk to have a Solvent Accessible Surface Distance (SASD) below 33 Å are approximately equal between all the methods. Although the length of a cross-link can be used to assess the quality of a model it, does not confirm that it is an accurate identification.

Table 6.2: Comparison of cross-link identification rates using xQuest, Jwalk and ValidateXL.py. As discussed in the text; a distance cut of of 33Å has been used in the Jwalk analysis. An xQuest score threshold of 20 has also been employed. The number of identified unique BSA cross-links has been compared to that remaining after validation by ValidateXL as described above.

Experiment	**Score Threshold**	**SASD <= 33 Å**	**ValidateXL**
DDA	131	46	39
IM-DDA Charged Stripped	128	48	24
HD-DDA Sample Calibrant	103	42	10

In Figure 6.7 the spectra for a cross-link with an xQuest linear discriminant score (LD-Score) of 26.87 and a SASD of 27.11 Å is shown. The fragmentation of the cross-link is very poor: there is only one annotated peak corresponding to the precursor m/z. None of the other peaks in the spectra can be matched to theoretical fragment ions for the candidate cross-link. Therefore despite the length of the cross-link and the xQuest score this is likely a false positive identification. ValidateXL rejects this cross-link assignment. The distance constraints provided by cross-link identifications do not account for the quality of fragmentation in a spectra. Whilst cross-link length can be used as a way to filter out models it is not suitable to test the validity of a cross-link:spectrum match. Cross-links must be assessed for fragmentation quality before distance restraints are calculated.

TCVADESHAGCEK-DTHKSEIAHR-a7-b2

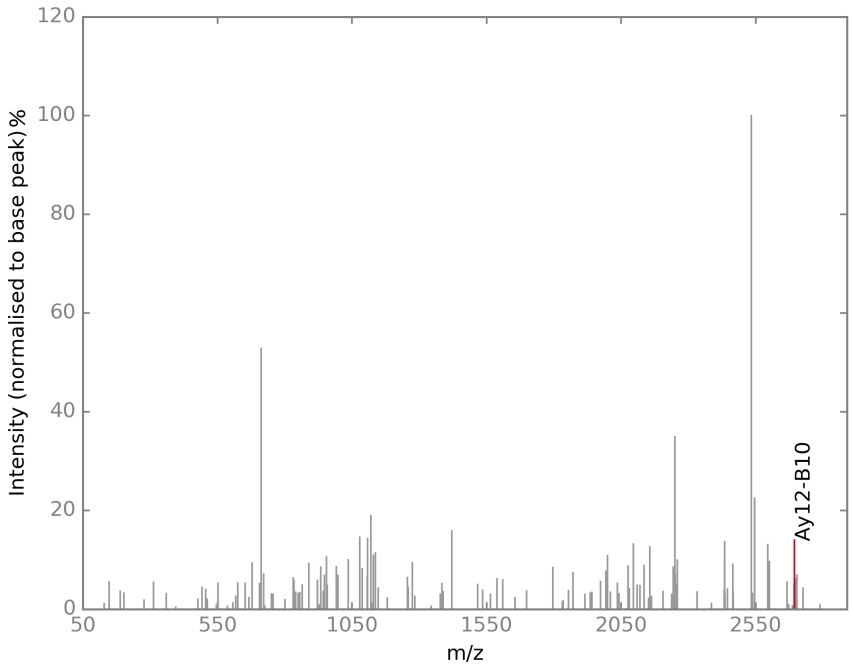


Figure 6.7: Example of a mis-assigned cross-link:spectrum match by xQuest from the analysis of cross-linked BSA using the Mid energy ramp. LD-score 26.87, SASD 27.11. Both the xQuest score and the SASD are within the suggested threshold for acceptance of the cross-link, however the spectral quality is very poor with only one annotated peak. ValidateXL rejects the cross-link.

In contrast to the results from the SASD analysis Table 6.2 reveals that the total number of cross-links passing manual validation from each experiment varies greatly. The DDA method outperforms both mobility methods. The HD-DDA method has the fewest validated cross-link:spectrum matches. As discussed in Section 5.4 there are two likely explanations for this: poor recombination of data during the merging of precursor data from the charge state families and poor synchronisation of the pusher in the ToF during data acquisition. In Chapter 3 the types of fragment ions missing from the xQuest search algorithms was discussed. A number of fragment ion types are not considered, in particular cross-linked ions where fragmentation events have occurred on both peptides (hereafter referred to as double fragmentation ions). It is possible that the number of cross-links in each set could be increased when these ion types are included in the analysis. While ValidateXL aids in filtering out poor cross-links it provides little insight into the SNR as it depends upon the xQuest annotations. In order to fully evaluate the effects of the mobility calibration in the

HD-DDA method a more in-depth analysis of the types of fragment ions and SNR is required. Hence we have developed AnnotateXL.

6.3.4 Annotate XL: Signal to Noise Improvement for QToF Experiments

In addition to the sequence coverage of both peptides assessed by ValidateXL, the SNR of a cross-link:spectrum match is a reliable measure of the accuracy of a cross-link assignment.[47] SNR is defined as the number of the matched peaks divided by number of observed peaks in an MS/MS spectra. AnnotateXL was developed to independently assess the quality of a cross-link:spectrum match. In addition to linear fragment ions and cross-link fragment ions where a single fragmentation event has occurred, AnnotateXL also consider the three ion types missing from the xQuest annotation: double fragmentation ions, immonium ions and diagnostic ions due to fragmentation of the amide bond in the BS3 cross-linker. To assess the SNR of all the cross-links identified by each of the QToF methods, AnnotateXL was executed over the full set of cross-links scoring in excess of twenty identified by xQuest.

To compare the performance of AnnotateXL to xQuest the SNR for each of the cross-links in the dataset was obtained according to the xQuest annotations and those calculated by AnnotateXL. As the HD-DDA Sample Calibrant experiment was not run in triplicate the DDA Mid ramp and IM-DDA Charged Stripped BSA experiments with the highest number of identifications were compared. To assess the relationship between SNR and cross-link size the comparison was carried out as a function of the number of amino acids in the cross-link. As expected AnnotateXL consistently reports a higher SNR than xQuest for each cross-link (Figure 6.8a). The largest increase in SNR can be seen in the DDA experiment where SNR increases by 0.4 (40%). The HD-DDA experiment shows the poorest increase in SNR. Even with the addition of previously unconsidered fragment ions types the SNR using AnnotateXL does not exceed 0.15 (15%). This indicates that the quality of the match between the cross-link and the spectra is poor. This is most likely due to inconsistent recombination of peaks from the various charge state families by MGFMerge.py (Section 5.4).

Table 6.3: Summary descriptive statistics for the difference between AnnotateXL and xQuest signal:noise ratio (SNR) for all tested QToF experimental methods. DDA is described in Chapter 3, IM-DDA with charge stripping is described in Chapter 4 and HD-DDA with BSA sample calibrant is described in Chapter 5. SNR for xQuest and AnnotateXL was calculated as described in the text.

Experiment	**SNR Mean Difference**	**SNR Std Difference**	**SNR Median Difference**	**SNR Min Difference**	**SNR Max Difference**
DDA	0.066	0.032	0.061	0.007	0.211
IM-DDA	0.062	0.027	0.061	0.008	0.115
HD-DDA	0.045	0.022	0.038	0.016	0.137

A high degree of variability is observed between the SNR for each cross-link. The values fluctuate widely across the range of the residue counts in each cross-link. To investigate this relationship further the difference in reported SNR between AnnotateXL and xQuest was plotted (Figure 6.8b). A moving average with a window length of 20 is also shown in green. For the all experimental methods this moving average can be seen to decrease with number of amino acids in the cross-link. This trend is most easily observed in the DDA experiment. As cross-link residue count increases the number of potential fragment ions generated also increases. This trend is thus expected. The mean difference in SNR is similar across each experiment: 0.066, 0.062 and 0.045 for the DDA, IM-DDA and HD-DDA experiments respectively (Table 6.3). The standard deviation is also observed to be similar. This indicates that AnnotateXL is increasing the annotations in a consistent manner.

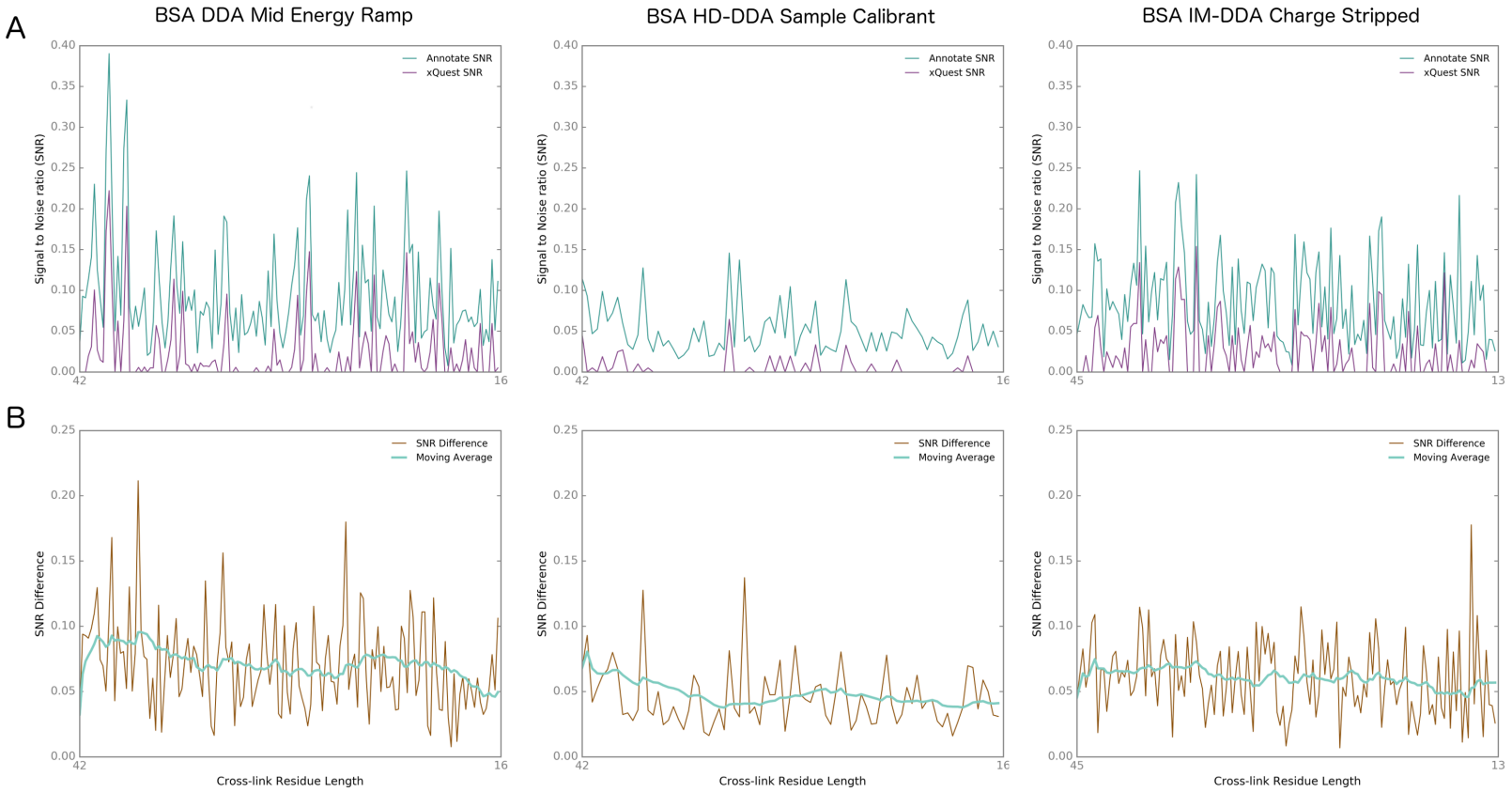


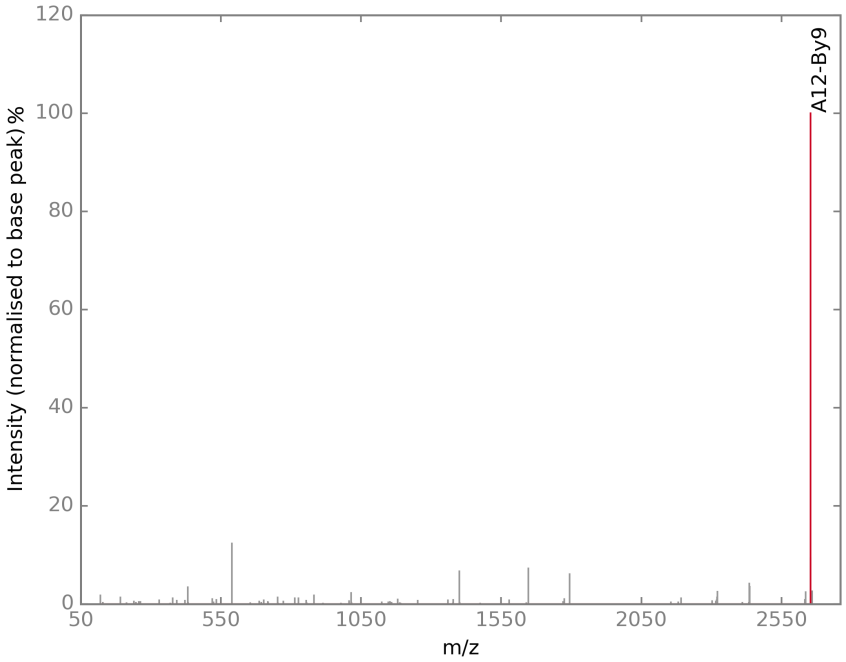
Figure 6.8: Signal to noise ratio (SNR) comparisons for all tested QToF cross-linking analysis methods, see Chapter 3 (DDA), Chapter 4 (IM-DDA with charge stripping) and Chapter 5 (HD-DDA with BSA sample calibrant) for more details. Data generated from the unique intra-molecular cross-links identified by xQuest analysis of the cross-linked BSA dataset. A) SNR as a function of cross-link residue length for spectra annotated by AnnotateXL (green) and xQuest (purple). B) SNR difference between xQuest and AnnotateXL (brown) and moving average (green) as a function of decreasing cross-link length.

The minimum improvement in annotation for the DDA experiment shows an increase of only 0.7%. The spectra for this cross-link:spectrum match is shown in Figure 6.9a. In this case only the precursor ion has been matched for the spectra, hence it is likely to be a mis-assignment and was categorised as a rejection by ValidateXL. Despite this poor level of annotation the cross-link has been assigned an LD score of 20.76.

The spectra for the largest SNR difference from the DDA experiment is shown in Figure 6.9b. This cross-link has an increase of 21% SNR compared to xQuest. This difference comes almost entirely from ions generated by fragmentation events on both the peptides in the cross-link. The tyrosine immonium ion can also be seen at 136 m/z. AnnotateXL accounts for both these ion types whilst xQuest does not.

Spectra for the two largest increases in SNR for the IM-DDA and HD-DDA experiments can be seen in Figures 6.10a and 6.10b. The cross-link:spectrum match shown in Figure 6.10a has a SNR difference of 11%, this is mostly composed of cross-linked fragment ions where both peptides have undergone fragmentation. The SNR when using AnnotateXL is 25% and for xQuest is 14%, hence it is likely to be correctly identified cross-link assignment. The crosslink:spectrum match in Figure 6.10b however, is most likely a mis-assignment. Although an increase in SNR when using AnnotateXL can be observed this is primarily due to the inclusion of immonium ions in the theoretical construction. xQuest matches only the Ab10 ion, that is, the b10 ion from the α peptide. The base peak in the spectra has not been matched and the SNR reported by AnnotateXL is only 13%.

a) EYEATLEECCAK-QTALVELLK-a5-b2, DDA lowest SNR increase



b) LAKEYEATLEECCAKDDPHACYSTVFDK-ALKAWSVAR-a3-b3, DDA highest SNR increase

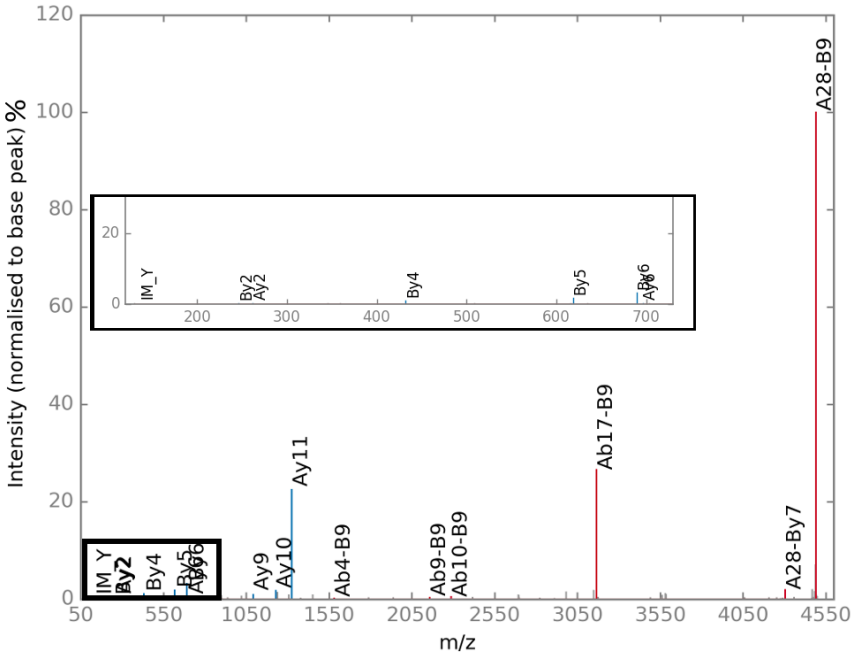
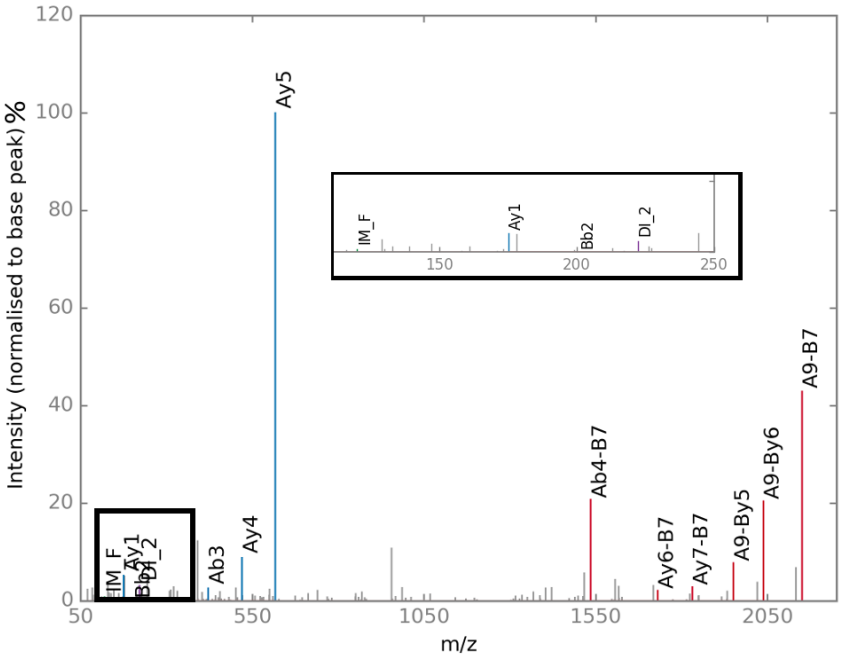


Figure 6.9: Cross-link:spectrum matches for the lowest (a) and the highest (b) SNR difference between AnnotateXL and xQuest in the DDA experiment. Cross-link sequence is displayed above each spectra. Annotated spectra were produced using an in house annotation script and MS/MS data for the cross-linked precursor from the original MGF files.

a) CCTKPESER-LSQKFPK-a4-b4, IM-DDA highest SNR increase



b) LFTFHADICTLPDTEKQIK-KQTALVELLK-a16-b3, HD-DDA highest SNR increase

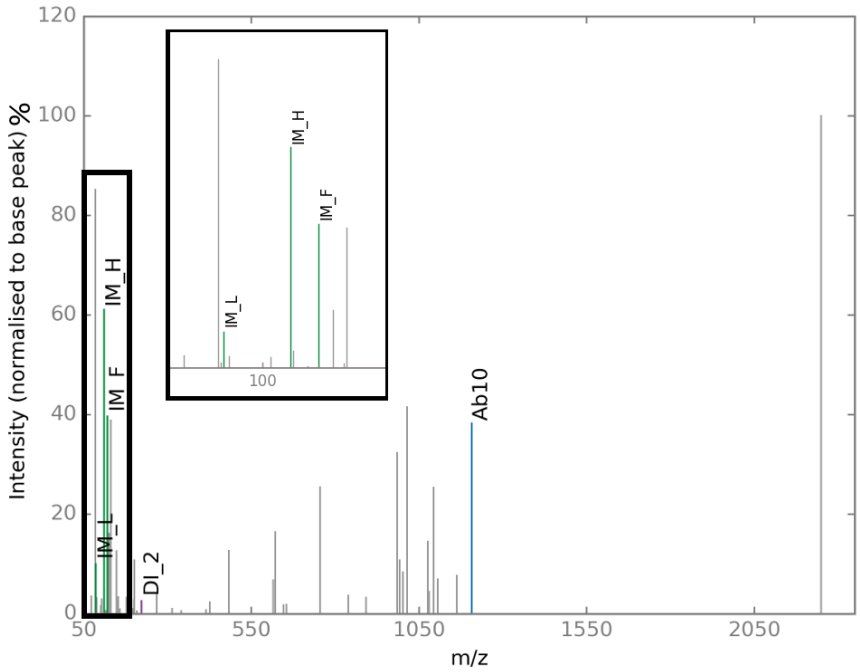


Figure 6.10: Cross-link:spectrum matches for the highest SNR difference between AnnotateXL and xQuest in the IM-DDA experiment (a) and the HD-DDA experiment (b). Cross-link sequence is displayed above each spectra. Annotated spectra were produced using an in house annotation script and MS/MS data for the cross-linked precursor from the original MGF files.

Both spectra in Figure 6.9b and 6.10a contain significant numbers of annotated cross-linked fragment ions where fragmentation has occurred on both peptides in the cross-link. As these ions are not expected by the xQuest annotation algorithms it is likely that they are responsible for the increase in SNR observed when using AnnotateXL. In order to evaluate the degree to which each fragment ion type increases the SNR, the mean of the difference in SNR for each ion type was calculated (Table 6.4). For both the DDA and IM-DDA experiments the greatest difference in fragment ions is observed for cross-linked fragment ions where a fragmentation event has occurred on both peptides. These fragment ions make a significant contribution to the increase in SNR. In all tested experiments the diagnostic ions offer the smallest contribute to the increase in SNR. As there are only two masses considered for the diagnostic ions this is expected.

Table 6.4: Mean difference between AnnotateXL and xQuest SNR by ion type for all tested QToF experimental methods. See Chapter 3 (DDA) Chapter 4 (IM-DDA with charge stripping) and Chapter 5 (HD-DDA with BSA sample calibrant) for more details on methods used. SNR was calculated for the all unique BSA intramolecular cross-links identified by xQuest. SNR was further broken-down based on the fragment ion type as determined by AnnotateXL.

Ion Type	**DDA Mean Difference**	**IM-DDA Mean Difference**	**HD-DDA Mean Difference**
Immonium ion	0.019	0.020	0.024
Diagnostic ion	0.006	0.004	0.008
Double fragmentation ion	0.041	0.037	0.013

For the HD-DDA experiment the immonium ions offer the largest increase in SNR, with double fragmentation ions yielding an increase of only 13%. The HD-DDA method uses calibration files to synchronise the pusher pulse to the arrival time of fragment ions from particular charge state families at the entrance of the orthogonal acceleration-Time of Flight analyser. In this way the method aims to increase the duty cycle of the instrument dynamically. As discussed in Section 5.3.4 there is a decrease in the sequence coverage of annotated

ions for the HD-DDA experiment. The calibration file used provides a single m/z value with which to synchronise the pusher pulse. As the SNR difference for double fragment ions is lower for the HD-DDA experiment this calibration may not represent the correct arrival time of these cross-linked fragment ions.

6.4 Conclusion and Further Work

xQuest is uniquely positioned in that it is offered with full source code and parallelisation of the analysis. This enables full understanding of the algorithm by the user and permits protein complexes to be analysed in short periods of time. Despite the inclusion of sophisticated scoring algorithms that combine to give a final linear discriminant score, there is no quantitative measure of sequence coverage for a cross-link:spectrum match. This information is available but must be accessed programatically from the XML result files.

ValidateXL was designed to extract this information and employ it to further evaluate cross-link identifications and help identify potential cross-link mis-assignments. By categorising cross-link identifications into three classes: *validated*, *undetermined* and *rejected*, ValidateXL greatly reduced the time required to manually validate cross-link identifications in each of the presented triplicate ramp tests. Following validation the Mid ramp remained the best performing energy ramp, identifying the highest quantity of cross-link identifications. The extraction of information relating to sequence coverage also permitted a more in-depth analysis of the fragmentation patterns in the ramps. In contrast to published results for the Orbitrap,[33,57] fragmentation of both peptides in the cross-link was observed to a higher degree of efficiency. As discussed in Section 6.3.3 this is most likely due to achieving more optimal energy ranges for both of the peptides in the cross-link rather than an optimal energy for the full m/z of the precursor.

An alternative method of assessing the quality a cross-link:spectrum match is by consideration of the SNR. AnnotateXL was developed to provide a measure of SNR using the given cross-link sequence and the MS/MS peak list. There are a number of fragment ion types which are not considered by xQuest that have been included in the AnnotateXL software. As expected AnnotateXL consistently determines a higher SNR for cross-link identification.

However, SNR increase alone is insufficient for a cross-link identification to be considered accurate. xQuest does not provide the user with a fully annotated spectra that includes all of the observed peaks. Instead an image is created containing peaks common to both the heavy and light versions of the cross-linked spectra. AnnotateXL creates a spectrum representation that includes all of the observed peaks, providing full annotations for those that have been matched.

Analysis of the HD-DDA experiments by AnnotateXL identified a reduction in the quantity of annotated cross-linked fragment ions. The reason for the significant reduction in validated cross-links for this method may be the result of the calibration used to synchronise the pusher pulse. Further investigation of the mobility of cross-linked fragment ions is needed to confirm this.

At present ions generated through neutral losses and internal cleavages have not been considered by AnnotateXL. It is also limited to deconvoluted MS/MS data that is singly charged. In addition, although the cross-linker can be easily changed by modifying the code a wider user base may not find this straightforward. In order for the project to become more viable to the wider community these changes would need to be implemented. Alternatively as the software is offered as open source under a GNU General Public License users with more computational expertise can modify the code base to suit their individual requirements.

Cross-linking mass spectrometry continues to gain popularity as a structural biology tool. Advancements in sample preparation and analysis methods are driving the technique towards a high-throughput workflow. The biggest bottleneck remains the evaluation of cross-link assignments. The standards of such assignments in publications has also been recently questioned.[47] The spectral quality of a cross-link:spectrum match is by far the best method of validating cross-link identifications. ValidateXL encourages this by design, directing manual inspection of cross-links to areas of largest uncertainty. AnnotateXL also encourages the user to consider the quality by assessing the SNR of a candidate cross-link:spectrum match by independent annotation.

Chapter 7

Conclusion

Structural biology aims to understand the function of the macromolecular machines that control vital cellular functions. As we continue to study larger and more complex multi-subunit assemblies the power of combining multiple structural techniques provides greater levels of insight. Over the last twenty years cross-linking mass spectrometry has become a valuable complementary tool for the structural biologist. One of the most recent examples of the power of such a combinatorial approach was the elucidation of the *Saccharomyces cerevisiae* nuclear pore complex. This study used a number of mass spectrometry approaches in combination with electron microscopy, small angle X-ray scattering and integrative modelling using atomic resolution structures previously collected by X-ray crystallography. Some 3077 unique cross-links residue pairs were identified for the 552 protein mega dalton complex. Combined with the additional structural techniques, these distance restraints aided the development of a model with the positions of all 552 nucleoporins (Nups) defined.

The continued evolution of cross-linker chemistries in addition to improvements in sample preparation techniques, MS analysis and data processing algorithms are responsible for the advancement of the technique. Such developments have been designed to improve coverage of the protein and also the rate of cross-link identification in MS data. It is these innovations that are driving demand for cross-linking as a complimentary structural biology tool.

Although much work has been accomplished in developing better tools and strategies, most cross-linking is carried out using Orbitrap analysers. As a result most software is written to expect Orbitrap style data. The protocol developed in Chapter 3 extends the method to

include QToF geometries, making cross-linking accessible to a wider community. The energy ramp tests reveal an optimal range of fragmentation energies that maintain cross-linked fragment ions, that have been previously observed to be absent from QToF analysis (Private communication). As the Synapt instrument (Waters Corp.) allows seamless integration of ion mobility into an experiment, establishment of this protocol enabled a study into the effects of coupling ion mobility separation to the analysis of cross-linked samples.

Following the commercialisation of Travelling Wave Ion Mobility ion mobility separation has been successfully used to separate charge state families as well as different biomolecules.[83,104] In Chapter 4 the effectiveness of ion mobility to isolate cross-linked precursors from linear peptides was evaluated. Although no increase in the rate of cross-link identification was observed, some separation of cross-linked precursors was achieved at high m/z. This could be further exploited in future studies by increasing cross-link length through means of a limited digest. Removal of singly charged precursor was also found to boost the cross-link yield of more complex protein samples. The rate of singly charge species reduction however, was observed to be limited.

The unique geometry of the Triwave stacked ring ion guide (SRIG) coupled with the application of a DC bias to either the trap or transfer SRIG also allows mobility separation to be carried out on fragment ions. A modified version of the HD-DDA method proposed by Helm et al. [44] was developed (Chapter 5). This method aimed to dynamically synchronise the pusher in the Time of Flight analyser (ToF) to increase the duty cycle of the instrument for fragment ions generated from cross-linked precursors. To generate this dynamic synchronisation, the mobility of all charge state families was required. This necessitates multiple analysis of the same sample and therefore greater sample volumes and analysis time. Despite trials of multiple calibrants no increase in the overall rate of cross-link identification was observed. Further refinement of the recombination parameters in the merge script may improve the overall outcome. However, the extra analysis time and sample requirements will likely limit the uptake of the method in the wider field.

Throughout the analysis xQuest was used to search for cross-links in the MS data. As one of the first pieces of cross-link software xQuest has been widely used by the community.[78,58,109,73,17,79] Only minor modifications were needed to enable the analysis of QToF

data with xQuest. Identification of these parameters proved challenging, with little documentation available. In particular the need for 32 bit encoding of the mzXML files was not well described and proved to be key to the evaluation process. The software application is offered under a GNU General Public License and as such is provided with full source code. This enables a comprehensive evaluation of the algorithms to ensure full understanding of the operating parameters. This avoids the dangers associated with "black box" algorithms where an understanding of their limitations and therefore the extent of their application, cannot be fully determined.

xQuest provides a suite of scoring algorithms to judge cross-link spectra on the basis of intensity, correlation and match probability. There are however, a number of parameters that are not considered. Iacobucci and Sinz [47] recommend that when validating assignment of a cross-link to an MS/MS spectra consideration should be given to the signal to noise ratio(SNR) and to the sequence coverage of both the peptides in the cross-link. In order to extract more useful information from the xQuest results and to reduce the time taken to manually assess cross-link assignments ValidateXL was developed. ValidateXL does not replace the need for manual validation. The ability to assess the quality of a cross-link assignment is an essential step in the evaluation of the experimental results. By filtering cross-links based on the sequence coverage of alpha peptide, beta peptide and cross-linked fragment ions the amount manual validation required was reduced by up to 50%. This resulted in a reduction in the overall identification rate of cross-links when compared to the use of a score threshold with the addition of raw data validation. The final cross-link assignments show an increase in SNR and sequence coverage. Consequently they were more reliable. For cross-linking to continue to be a valued contribution to structural modelling approaches the risk of false positive assignments being included in literature must be minimised.

Innovations in cross-linking analysis are advancing the technique towards a more high throughput approach that continues to be in high demand. This growth requires the development of a set of guidelines to ensure a high standard of published data. We have shown that the application of a score threshold to validate cross-links is necessary but not sufficient to distinguish genuine assignments from mis-assigned identifications. Furthermore, the current standard of published results is quite broad with some instances failing to in-

clude cross-link scores or amino acid distances and most frequently not providing spectra. The need for improvement is clear. The COST BM1403 initiative cross-linking work group (*http://structuralproteomics.eu/*) has begun to address this need with a series of workshops conducted at annual structural proteomics symposiums. This active approach along with continued improvements to the experimental preparation, mass spectrometry analysis, computational solutions and validation will allow cross-linking mass spectrometry to reach full maturity.

Appendix A: xQuest Ubuntu Installation Protocol

Installing xQuest onto Ubuntu 14.04

1) Download install_xquest.sh from

http://proteomics.ethz.ch/cgi-bin/xquest2_cgi/installation.cgi and copy it to a new directory. This is the directory where xQuest will be located. Change the permissions of the file to make it executable.

cd home

mkdir xquest

cd xquest

cp home/Downloads/install_xquest.sh .

chmod +x install_xquest.sh

To install required packages: *./install_xquest*

Say Yes to all prompts in the command line and use defaults

2) Download and unzip xQuest/xProphet from

http://proteomics.ethz.ch/cgi-bin/xquest2_cgi/download.cgi

Copy and unzip the xQuest zip file to the directory

cp home/Downloads/V2_1_1.zip .

unzip V2_1_1.zip

3) Change to the installation folder and edit the first line in the install_xquest.sh script change INSTALLDIR=/home/xqxp/xquest/V2_1_1/xquest to the path you have just created when unzipping xQuest, save the file. Check you perl installation path:

which perl

If it is not /usr/bin/perl you will need to amend the xQuest changeheader.pl script. Open the file in a text editor

Change the top line from usr/bin/perl to your perl installation

4) Add the xquest/bin directory to your PATH

cd home

emacs .bash_profile

Add the following to the file

export PATH=/your_home_directory/xquest/V2_1_1/xquest/bin:$PATH

Run the following command to execute the new link

source home/.bash_profile

Installing the web server

1) Create a softlink to the CGI bin directory

cd usr/lib/cgi-bin

sudo ln -s /your_home_directory/xquest/V2_1_1/xquest/cgi xquest

2) Change to Apache available sites

cd /etc/apache2/sites-available

3) Copy default configuration file to a new file called xquest.conf

sudo cp 000-default.conf xquest.conf

chmod 777 xquest.conf

4) Deactivate the default and activate the new configuration file

sudo a2dissite 000-default.conf

sudo a2ensite xquest.conf

5) Reload apache2

sudo service apache2 reload

6) Check for mod_perl, mod_alias and mod_cgi or mod_cgid in /etc/apache2/mods-enabled

cd /etc/apache2/mods-enabled ls -l

If they are not present check for them in /etc/apache2/mods-available, if found enable them:

a2enmod perl

a2enmod cgi

a2enmod alias

If they are not in mods-available:

sudo apt-get install libapache2 mod-alias

sudo apt-get install libapache2 mod-perl

apt-get install libapache2 mod-cgi

7) Change permissions on the cgi

cd home/xquest/V2_1_1/xquest/

sudo chown -R root:root cgi

8) Reload apache2

sudo service apache2 reload

9) Create a soft link so localhost can see the results folders

Create a directory for xquest results. Be sure to create a new directory and not to use the results folder in the installation directory that comes with xquest

cd home

mkdir xquext_results

Create a soft link from the var directory to the new results folder

cd /var/www

sudo ln -s /your_home_directory/xquest_results results

10) Back up and open Environment.pm and change the hostname

The output in the command line is your hostname, this needs to be entered into environment.pm

Edit the Environment.pm file

cd home/xquest/V2_1_1/xquest/modules/

cp Environment.pl Environment.pm.bak

emacs Environment.pm

Change the lines as follows:

$machines'your_local_host' = "your_local_host";

...

/your_home_directory/V2_1_1/xquest

...

$serverpaths('your_local_host')('xquest_stable') = "/your_home_directory/V2_1_1/xquest";

$serverpaths('your_local_host'()'web.config') =

"/your_home_directory/V2_1_1/xquest/conf/web.config";

$serverpaths('your_local_host')('mass.def') = "/your_home_directory/V2_1_1/xquest/deffiles/mass_table.def";

11) Back up and edit webconfig file to point to the results folder and the soft link created in www/var/

cd home/xquest/V2_1_1/xquest/conf/

cp web.config web.config.bak

emacs web.config

Change the following lines to point to the result folder and to the softlink

resultdirbase::/your_home_directory/xquest/xquest_results

resulturlbase::http://localhost/results

12) Reload apache2

sudo service apache2 reload

xQuest should now be successfully installed

13) Create an inputs directory for the searches in xQuest

mkdir home/xquest_inputs

Amend the broken soft link

1) Open the pQuest.pl script

cd home/xquest/V2_1_1/xquest/bin/

gedit pQuest.pl

2) On line 154 change

my $cmd = "ln -s $centroidmzxmlfile $basename/$mzxmlfilename";

print(–>$cmd \n);

to

my $cmd = "ln -s ../$centroidmzxmlfile $basename/$mzxmlfilename";

print(–>$cmd \n);

3) Save and close pQuest.pl

Appendix B: xQuest Search Parameters

Enzyme and Peptide Settings:

Parameter	Value	Description
missed-cleavages	2	Maximal missed cleavages
mindigestlength	5	Minimal peptide size in AA
maxdigestlength	50	Maximal peptide size in AA
nocutatxlink	1	If set the enzyme will not cut at the crosslinkerposition
variable-mod	M:1599491	Variable modification example: M:15.99491 one variable modification can be defined
nvariable-mod	1	Number of variable modifications per peptide
ionseries	010010	abcxyz consider ionseries (1)
ioncharge-common	1	Charge for common ions
ioncharge-xlink	1	Charge for xlink ions
xlinktypes	1111	Defines types of ions to search for in enumeration mode monolink:intralink:intraprotein-x-link:interprotein-x-link (only applicable in enumeration mode)
ntermxlinkable	1	If (1) modifies the first peptide of a protein add then also amino acid Z to AArequired

Crosslinker Settings:

Parameter	Value	Description
AArequired	K,S,T,Y,Z	Aminoacid that is cross-linked for more than one AA indicate AAs separated by comma
xkinkerID	BS3	Xlinker name
xlinkermw	138.0680796	Mass shift for intra or inter-peptide cross-links
monolinkmw	156.0786443 155.0946287	Mass shift for monolinked peptides separated by comma

Output Settings:

Parameter	Value	Description
drawspectra	1	Draw spectrum-plots

Fixed Modifications:

Parameter	Value
C	57.02146

MS Settings:

Parameter	Value	Description
tolerancemeasure	ppm	Da or ppm
ms1tolerance	5	Tolerance for precursor mass matching MS1
ms1tol-minborder	-10	Defines asymmetric borders for matching; tolerances unit is tolerance measure
ms1tol-maxborder	10	Defines asymmetric borders for matching; tolerances unit is tolerance measure
tolerancemeasure-ms2	ppm	Da or ppm
ms2tolerance	10	Tolerance for peak matching on MS2 for linear ions
xlink-ms2tolerance	10	Tolerance for MS2 matching for cross-linked ions
minionsize	200	Minimum ion size in MS2 mode to be considered
maxionsize	2000	Maximum ion size in MS2 mode to be considered

Appendix C: Kernel Density Estimation

7.1 Kernel Density Estimation

In Section 3.3.4 Kernel Density Estimation (KDE) has been used to estimate the true underlying distribution of the data. KDE creates a kernel function at every datum with the point at its centre. The underlying probability density function (PDF) for the data is estimated by summing these kernel functions and dividing by the number of points in the data. This ensures that the definite integral of the PDF is 1 and the values are non-negative. KDE is defined by Equation 7.1 where n is the number of points in the data, K is the chosen kernel function and h is the band width of smoothing parameter.

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{7.1}$$

The most complicated parameter to optimise for KDE is the bandwidth.[19] This parameter describes the standard deviation of the kernel function applied to each data point. If the value is small most of the probability density is placed on the data point, if it is large the probability density is spread out to the neighbouring data points. If the selected parameter is too large the KDE is over-smoothed and may ignore key features of the data, if it is too small the KDE appears more volatile. Much research has been done into the optimisation of the bandwidth parameter. As we are estimating univariate data with a gaussian kernel function we have implemented Silverman's rule of thumb to calculate the bandwidth that minimise the mean integrated squared error.[39] Silverman's rule of thumb defines h to be:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}} \tag{7.2}$$

In order to use KDE data must be independent and identically distributed. These attributes are present in data collected for cross-link subscores. As the subscores are based upon MS/MS spectra and the method of fragmentation is constant throughout the experiment cross-link score data is identically distributed. Furthermore the probability of the score that is assigned to a cross-link does not depend on the score assigned to any other. The data are also therefore independent.

7.2 Appendix D: BSA Peptides with a Charge State above +3

Table 7.1: Linear peptides identified with charge states of +4 and +5

Sequence	Precursor MH+ (Da)	z	MH+ Error (ppm)	Retention Time (min)	Peptide Type	Drift (bins)
(K)SHCIAEVEKDAIPENLPPLTADFAEDKDVCK(N)	3511.67	5	2.21	58.76	Missed Cleavage	88.26
(K)SHCIAEVEKDAIPENLPPLTADFAEDKDVCK(N)	3511.67	5	-0.41	57.88	Missed Cleavage	86.86
(K)SHCIAEVEKDAIPENLPPLTADFAEDKDVCK(N)	3511.66	5	-1.81	55.90	Missed Cleavage	86.42
(R)LAKEYEATLEECCAKDDPHACYSTVFDK(L)	3350.47	4	3.67	45.77	Missed Cleavage	79.14
(R)LAKEYEATLEECCAKDDPHACYSTVFDK(L)	3350.47	4	2.72	50.16	Missed Cleavage	79.24
(K)TVMENFVAFVDKCCAADDKEACFAVEGPK(L)	3324.47	4	2.44	58.79	Variable Modification	80.48
(K)TVMENFVAFVDKCCAADDKEACFAVEGPK(L)	3324.46	4	-0.92	56.15	Variable Modification	79.78
(K)TVMENFVAFVDKCCAADDKEACFAVEGPK(L)	3308.47	4	0.56	65.71	Missed Cleavage	80.05
(K)TVMENFVAFVDKCCAADDKEACFAVEGPK(L)	3308.47	4	0.23	73.35	Missed Cleavage	80.77

Appendix E: Further Methods for ValidateXL and AnnotateXL

Further Methods for ValidateXL

Calculation of Theoretical Ions

Figure 7.1: Fragmentation of a cross-link and ions generated. Cross-linked ions shown in red, linear ions in green. Position of the cross-linker shown by red line, linked amino acids highlighted in red

In order to generate the sequence coverage for each peptide the theoretical number of fragment ions is first calculated using the length of each peptide. For cross-linked peptides both linear and cross-linked fragment ions must be considered (Figure 7.1). For linear ions Equation 7.3 shows the calculation of the number of theoretical linear fragment ions.

$$n(X_{\text{pos}} - 2) + c(L_{\text{pep}} - X_{\text{pos}}) \tag{7.3}$$

where n is the number of N terminal series ions you wish to calculate (a, b or c) and c is the number of C terminal series ions you wish to calculate (x, y or z). L_{pep} is the length of the alpha or beta peptide and X_{pos} is the position of cross-linked amino acid in the peptide. In the first term two is subtracted to accommodate the absence of a b1 ion. As discussed in Section 1.5.1 these cannot form without an additional carbonyl necessary to generate the oxazolone structure. For the b and y series only Equation 7.3 can be reduced to $L_{\text{pep}} - 2$.

For cross-linked fragment ions the total number of theoretical fragment ions for a cross-link is calculated according to Equation 7.4, where the variables are defined as above. For the b and y ion series only this can be reduced to $L_{\text{pep}} - 1$.

$$n(L_{\text{pep}} - X_{\text{pos}}) + c(X_{\text{pos}} - 1) \tag{7.4}$$

Consideration of Sequence Coverage Threshold

As ValidateXL is designed for use with xQuest/xProphet application this calculation method approximates the fragment ions that xQuest considers. As such it does not include ions generated by double fragmentation events of cross-linked peptides, immonium ions, neutral losses, internal cleavage ions or diagnostic cross-linker ions. Following calculation of the theoretical ions the final sequence coverage calculated by ValidateXL is simply the ratio of matched ions to theoretical ions.

ATEEQLKTCMENFVAFVDK-KQTALVELLK-a7-b1 Score 37

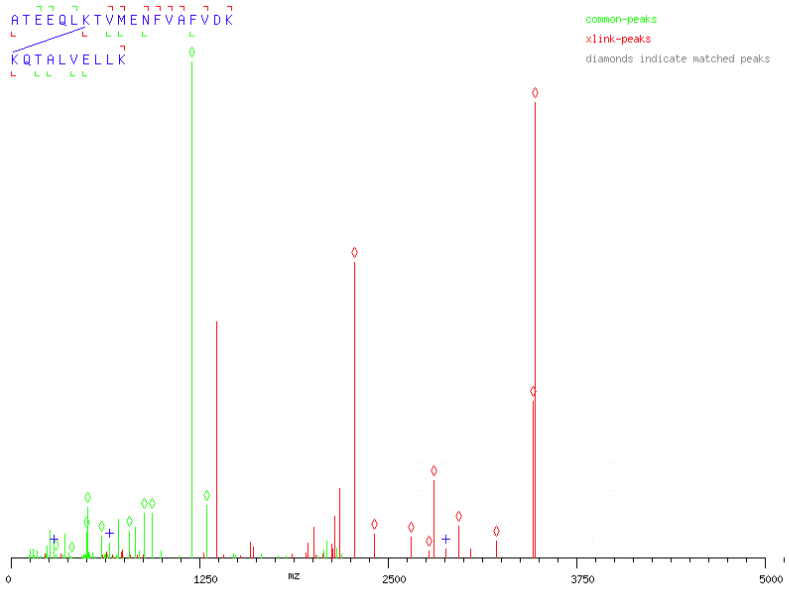
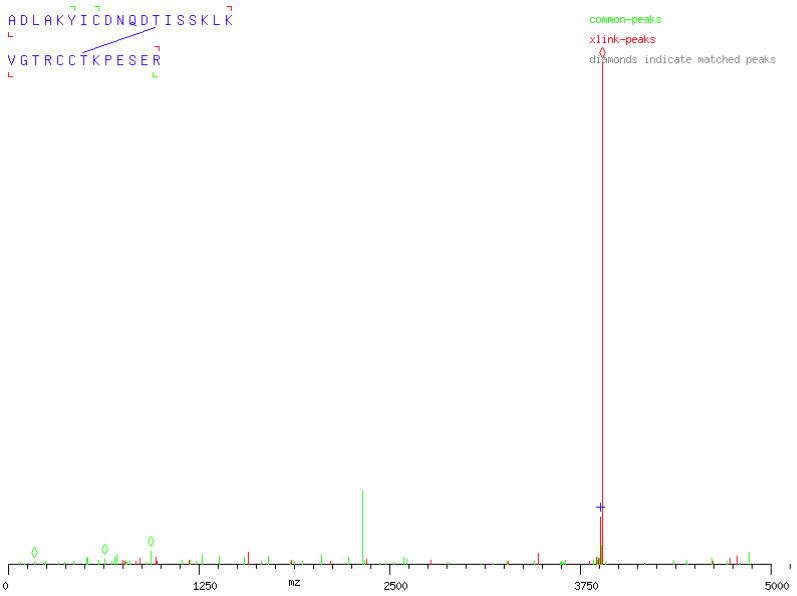


Figure 7.2: Cross-link filtered out by ValidateXL when using sequence coverage of 40% for linear and cross-linked fragment ions

a) ADLAKYICDNQDTISSKLK-VGTRCCTKPESER-a13-b7, Score 21



b) FWGKYLYEIARR-IETXREK-a4-b3, Score 27

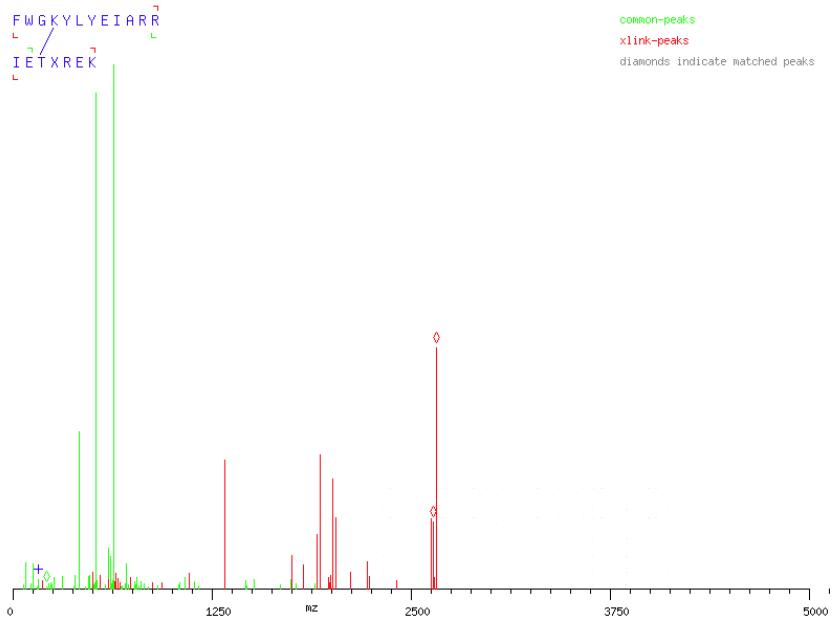


Figure 7.3: Cross-link mis-assignments filtered out by ValidateXL but included in xQuest result when using a score threshold with raw data validation

As discussed section 1.6.3 (Introduction) cross-link data has no ground truth. It is not possible to confirm with 100% certainty that a spectra contains a cross-linked peptide and that the identification assigned to the spectra is the correct one. As such an investigation

into the sensitivity and specificity (Receiver Operating Characteristics) has not been carried out. To determine the threshold for sequence coverage used in categorising cross-links by ValidatedXL, cross-link:spectrum matches from the best performing DDA ramp were manually evaluated. Following manual validation of the Mid ramp DDA dataset a sequence coverage of 30% was selected as a threshold for both linear and cross-linked fragment ions. At values above 30% cross-link:spectrum matches as demonstrated in Figure 7.2 were found to be classified as requiring further validation. A sequence coverage of $\leq 30\%$ was calculated for the alpha peptide linear fragment ions. In this case the length of the alpha peptide and position of the crosslinker limit the number of linear fragment ions that can be produced. Figure 7.3a show two examples of cross-links that were omitted from the validated results but included in the original dataset. The spectra shown in Figure 7.3a is the result of incomplete precursor fragmentation. The base peak in the spectra represents the intact precursor ion. Figure 7.3b is a mis-assigned cross-link. The only annotated peaks in the spectra are for the intact precursor, the b2 fragment ion of the β peptide and y1 ion of the α peptide. The SNR for this spectra is too low to consider the cross-link:spectrum match as accurate.

Further Methods for AnnotateXL

AnnotateXL Nomenclature

As cross-linked precursors contain two peptides an adaptation to the Roepstorff nomenclature is required to annotated the range of peaks that can be created during fragmentation. In order to accurately label peaks in the MS/MS spectra the following nomenclature was used. Linear fragment ions use the Roepstorff nomenclature defined by Roepstorff and Fohlman [88] with a prefix to identify the peptide. For simplicity the α peptide is labelled A and the β peptide is labelled B. For example Ab2 refers to the b2 ion from the α peptide. Cross-linked fragment ions where two fragmentation events have occurred, one on each peptide, follow a similar pattern. For example Ab2-By3 refers to the b2 ion from the α peptide cross-linked to the y3 ion from the β peptide. For cross-linked peptides where fragmentation has occurred on only one peptide the complete peptide is numbered based on the total number of amino acid residues in that peptide. For example Ab2-y9 refers to the α b2 ion cross-linked to the

complete β peptide.

AnnotateXL Creation of Fragment Ions

The generation of theoretical fragment ions in AnnotateXL is handled by the `Fragmenter` class. Generation of the ions is carried out in a series of steps. These are represented below for the cross-link displayed in Figure 7.4:

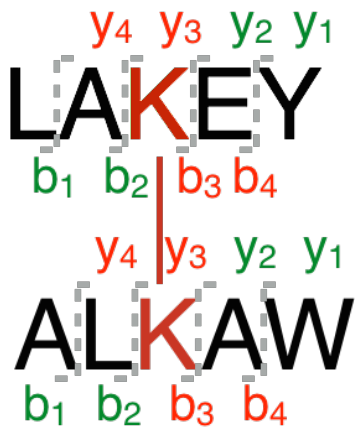


Figure 7.4: Schematic of cross-link nomenclature and theoretical fragment ion generation

Step1:

The N terminal fragment ions are generated for a peptide by sequentially looping over the amino acids in each peptide sequence for example "LAKEY" becomes: "L"

"LA"

"LAK"

"LAKE"

"LAKEY"

For the C terminal ions the peptide sequence is reversed to give: "Y"

"YE"

"YEK"

"YEKA"

"YEKAL"

Step 2:

Once the sequences have been correctly formed the theoretical linear fragment ions are calculated by enumerating across the strings up to the amino acid before the position of the cross-linker on each peptide. For the α peptide shown in Figure 7.4 these are: "L"
"LA"
"Y"
"YE"

Step 3:

For cross-linked ions the correct set of amino acid sequences is generated by subtracting the set of theoretical linear fragment ions from the full set generated in Step 1. To avoid duplication the N and C terminal ions are generated separately and last string is only generated if the peptide is complete. For the α peptide in Figure 7.4 these are: Linked N' terminal series:
"LAK"
"LAKE"
"LAKEY"
Linked C' terminal series:
"YEK"
"YEKA"
"YEKAL"

Step 4:

In order to correctly calculate the mass of the theoretical cross-linked ions, single and double fragmentations events are generated in separate groups. Complete peptides require N and C terminal modifications representing two hydrogen and one oxygen atom. These mass calculations are carried out by the `xl_fragment_ion` subclass. To generate the correct string series a cartesian product is used. For single fragmentation events the cartesian product of the linked ion series for the α peptide with the complete β peptide and the cartesian product

of the linked β peptide with the complete α peptide represent the full set of theoretical ions. For double fragmentation events the cartesian product of each set of linked fragment ions is generated. For the cross-link shown in Figure 7.4 all possible cross-linked fragment ions and the correct nomenclature are shown in Table 7.2.

Step 5:

After the correct sequence strings are created the `Fragmenter` class inherits from each ion type subclass in order to generate the correct mass and nomenclature for each of the theoretical fragment ions.

Table 7.2: Theoretical cross-linked fragment ions for cross-link in Figure 7.4

Sequence	**Nomenclature**	**Ion Description**	**Fragmentation Event**
LAK-ALKAW	Ab3-B5	α N terminal linked β Complete	Single
LAKE-ALKAW	Ab4-B5	α N terminal linked β Complete	Single
KEY-ALKAW	Ay3-B5	α C terminal linked β Complete	Single
AKEY-ALKAW	Ay4-B5	α C terminal linked β Complete	Single
LAKEY-ALK	A5-Bb3	β N terminal linked α Complete	Single
LAKEY-ALKA	A5-Bb4	β N terminal linked α Complete	Single
LAKEY-KAW	A5-By3	β C terminal linked α Complete	Single
LAKEY-LKAW	A5-By4	β C terminal linked α Complete	Single
LAK-ALK	Ab3-Bb3	α N terminal linked β N terminal linked	Double
LAK-ALKA	Ab3-Bb4	α N terminal linked β N terminal linked	Double
LAKE-ALK	Ab4-Bb3	α N terminal linked β N terminal linked	Double
LAKE-ALKA	Ab4-Bb4	α N terminal linked β N terminal linked	Double
LAK-KAW	Ab3-By3	α N terminal linked β C terminal linked	Double
LAK-LKAW	Ab3-By4	α N terminal linked β C terminal linked	Double
LAKE-KAW	Ab4-By3	α N terminal linked β C terminal linked	Double
LAKE-LKAW	Ab4-By4	α N terminal linked β C terminal linked	Double
KEY-ALK	Ay3-Bb3	α C terminal linked β N terminal linked	Double
KEY-ALKA	Ay3-Bb4	α C terminal linked β N terminal linked	Double

AKEY-ALK	Ay4-Bb3	α C terminal linked β N terminal linked	Double
AKEY-ALKA	Ay4-Bb4	α C terminal linked β N terminal linked	Double
KEY-KAW	Ay3-By3	α C terminal linked β C terminal linked	Double
KEY-LKAW	Ay3-By4	α C terminal linked β C terminal linked	Double
AKEY-KAW	Ay4-By3	α C terminal linked β C terminal linked	Double
AKEY-LKAW	Ay4-By4	α C terminal linked β C terminal linked	Double

Bibliography

[1] Frank Alber, Svetlana Dokudovskaya, Liesbeth M Veenhoff, Wenzhu Zhang, Julia Kipper, Damien Devos, Adisetyantari Suprapto, Orit Karni-Schmidt, Rosemary Williams, Brian T Chait, et al. The molecular architecture of the nuclear pore complex. *Nature* 450.7170 (2007), 695–701.

[2] Alexander von Appen, Jan Kosinski, Lenore Sparks, Alessandro Ori, Amanda L DiGuilio, Benjamin Vollmer, Marie-Therese Mackmull, Niccolo Banterle, Luca Parca, Panagiotis Kastritis, et al. In situ structural analysis of the human nuclear pore complex. *Nature* 526.7571 (2015), 140.

[3] Michael Barber, Robert S Bordoli, Gerard J Elliott, R Donald Sedgwick, Andrew N Tyler, and Brian N Green. Fast atom bombardment mass spectrometry of bovine insulin and other large peptides. *Journal of the Chemical Society, Chemical Communications* 16 (1982), 936–938.

[4] Michael Barber, Robert S. Bordoli, R. Donald Sedgwick, and Andrew N. Tyler. Fast atom bombardment of solids (F.A.B.): a new ion source for mass spectrometry. *J. Chem. Soc., Chem. Commun.* (7 1981), 325–327.

[5] Sean A Beausoleil, Mark Jedrychowski, Daniel Schwartz, Joshua E Elias, Judit Villén, Jiaxu Li, Martin A Cohn, Lewis C Cantley, and Steven P Gygi. Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proceedings of the National Academy of Sciences of the United States of America* 101.33 (2004), 12130–12135.

[6] Scarlet Beck, Annette Michalski, Oliver Raether, Markus Lubeck, Stephanie Kaspar, Niels Goedecke, Carsten Baessmann, Daniel Hornburg, Florian Meier, Igor Paron, et

al. The Impact II, a very high-resolution quadrupole time-of-flight instrument (QTOF) for deep shotgun proteomics. *Molecular & Cellular Proteomics* 14.7 (2015), 2014–2029.

[7] Adam Belsom, Michael Schneider, Lutz Fischer, Oliver Brock, and Juri Rappsilber. Serum albumin domain structures in human blood serum by mass spectrometry and computational biology. *Molecular & Cellular Proteomics* (2015), mcp–M115.

[8] Klaus Biemann. Contributions of mass spectrometry to peptide and protein structure. *Biological Mass Spectrometry* 16.1-12 (1988), 99–111.

[9] Klaus Biemann. *Mass spectrometry: organic chemical applications*. McGraw-Hill, 1962.

[10] Michael T Bowers, Paul R Kemper, Gert von Helden, and Petra AM van Koppen. Gas-phase ion chromatography: transition metal state selection and carbon cluster formation. *Science* 260.5113 (1993), 1446–1451.

[11] Robert Boyd and Árpád Somogyi. The mobile proton hypothesis in fragmentation of protonated peptides: a perspective. *Journal of the American Society for Mass Spectrometry* 21.8 (2010), 1275–1278.

[12] Nicholas I Brodie, Konstantin I Popov, Evgeniy V Petrotchenko, Nikolay V Dokholyan, and Christoph H Borchers. Solving protein structures using short-distance cross-linking constraints as a guide for discrete molecular dynamics simulations. *Science advances* 3.7 (2017), e1700479.

[13] Joshua Matthew Allen Bullock, Jannik Schwab, Konstantinos Thalassinos, and Maya Topf. The importance of non-accessible crosslinks and solvent accessible surface distance in modeling proteins with restraints from crosslinking mass spectrometry. *Molecular & Cellular Proteomics* 15.7 (2016), 2491–2500.

[14] Julia Maria Burkhart, Cornelia Schumbrutzki, Stefanie Wortelkamp, Albert Sickmann, and René Peiman Zahedi. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *Journal of proteomics* 75.4 (2012), 1454–1462.

[15] IA Buryakov, EV Krylov, EG Nazarov, and U Kh Rasulev. A new method of separation of multi-atomic ions by mobility at atmospheric pressure using a high-frequency amplitude-asymmetric strong electric field. *International Journal of Mass Spectrometry and Ion Processes* 128.3 (1993), 143–148.

[16] Murat A Cevher, Yi Shi, Dan Li, Brian T Chait, Sohail Malik, and Robert G Roeder. Reconstitution of active human core Mediator complex reveals a critical role of the MED14 subunit. *Nature structural & molecular biology* 21.12 (2014), 1028.

[17] William CH Chao, Yasuto Murayama, Sofía Muñoz, Andrew W Jones, Benjamin O Wade, Andrew G Purkiss, Xiao-Wen Hu, Aaron Borg, Ambrosius P Snijders, Frank Uhlmann, et al. Structure of the cohesin loader Scc2. *Nature communications* 8 (2017), 13952.

[18] Juan D Chavez, Chi Fung Lee, Arianne Caudal, Andrew Keller, Rong Tian, and James E Bruce. Chemical crosslinking mass spectrometry analysis of protein conformations and supercomplexes in heart tissue. *Cell systems* 6.1 (2018), 136–141.

[19] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology* 1.1 (2017), 161–187.

[20] David E Clemmer and Martin F Jarrold. Ion mobility measurements and their applications to clusters and biomolecules. *Journal of Mass Spectrometry* 32.6 (1997), 577–592.

[21] Martin J Cohen and FW Karasek. Plasma chromatography a new dimension for gas chromatography and mass spectrometry. *Journal of Chromatographic science* 8.6 (1970), 330–337.

[22] Loredana Lo Conte, Cyrus Chothia, and Joël Janin. The atomic structure of protein-protein recognition sites. *Journal of molecular biology* 285.5 (1999), 2177–2198.

[23] JHJ Dawson and M Guilhaus. Orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Communications in Mass Spectrometry* 3.5 (1989), 155–159.

[24] E De Hoffmann and V Stroobant. *Mass spectrometry: principles and applications*. Vol. 8. Wiley, 2007.

[25] Shannon Eliuk and Alexander Makarov. Evolution of Orbitrap mass spectrometry instrumentation. *Annual Review of Analytical Chemistry* 8 (2015), 61–80.

[26] Lord Rayleigh F.R.S. XX. On the equilibrium of liquid conducting masses charged with electricity. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 14.87 (1882), 184–186.

[27] John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246.4926 (1989), 64–71.

[28] Larissa S Fenn, Michal Kliman, Ablatt Mahsut, Sophie R Zhao, and John A McLean. Characterizing ion mobility-mass spectrometry conformation space for the analysis of complex biological samples. *Analytical and bioanalytical chemistry* 394.1 (2009), 235–244.

[29] Javier Fernandez-Martinez, Seung Joong Kim, Yi Shi, Paula Upla, Riccardo Pellarin, Michael Gagnon, Ilan E Chemmama, Junjie Wang, Ilona Nudelman, Wenzhu Zhang, et al. Structure and function of the nuclear pore complex cytoplasmic mRNA export platform. *Cell* 167.5 (2016), 1215–1228.

[30] Christian K Frese, AF Maarten Altelaar, Marco L Hennrich, Dirk Nolting, Martin Zeller, Jens Griep-Raming, Albert JR Heck, and Shabaz Mohammed. Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. *Journal of proteome research* 10.5 (2011), 2377–2388.

[31] Scott J Geromanos, Johannes PC Vissers, Jeffrey C Silva, Craig A Dorschel, Guo-Zhong Li, Marc V Gorenstein, Robert H Bateman, and James I Langridge. The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics* 9.6 (2009), 1683–1695.

[32] Sven H Giese, Adam Belsom, and Juri Rappsilber. Optimized Fragmentation Regime for Diazirine Photo-Cross-Linked Peptides. *Analytical chemistry* 88.16 (2016), 8239–8247.

[33] Sven H Giese, Lutz Fischer, and Juri Rappsilber. A study into the CID behavior of cross-linked peptides. *Molecular & Cellular Proteomics* (2015), mcp–M115.

[34] K Giles, JL Wildgoose, SD Pringle, and RH Bateman. Utilising Ion Mobility Spectrometry to separate precursors from background ions and species of different charges in automated tandem MS experiments. *Proc. 52nd ASMS Conf. Mass Spectrometry and Allied Topics*. 2004.

[35] Kevin Giles, Jonathan P Williams, and Iain Campuzano. Enhancements in travelling wave ion mobility resolution. *Rapid Communications in Mass Spectrometry* 25.11 (2011), 1559–1566.

[36] Kevin Giles, Jason L Wildgoose, David J Langridge, and Iain Campuzano. A method for direct measurement of ion mobilities using a travelling wave ion guide. *International Journal of Mass Spectrometry* 298.1-3 (2010), 10–16.

[37] Kevin Giles, Jason L Wildgoose, David J Langridge, and Iain Campuzano. A method for direct measurement of ion mobilities using a travelling wave ion guide. *International Journal of Mass Spectrometry* 298.1-3 (2010), 10–16.

[38] Kevin Giles, Steven D Pringle, Kenneth R Worthington, David Little, Jason L Wildgoose, and Robert H Bateman. Applications of a travelling wave-based radio-frequency-only stacked ring ion guide. *Rapid Communications in Mass Spectrometry* 18.20 (2004), 2401–2414.

[39] Michael Götze, Jens Pettelkau, Romy Fritzsche, Christian H Ihling, Mathias Schäfer, and Andrea Sinz. Automated assignment of MS/MS cleavable cross-links in protein 3D-structure analysis. *Journal of The American Society for Mass Spectrometry* 26.1 (2015), 83–97.

[40] Michael Götze, Jens Pettelkau, Sabine Schaks, Konstanze Bosse, Christian H Ihling, Fabian Krauth, Romy Fritzsche, Uwe Kühn, and Andrea Sinz. StavroX a software for analyzing crosslinked products in protein interaction studies. *Journal of the American Society for Mass Spectrometry* 23.1 (2012), 76–87.

[41] John Greaves and John Roboz. *Mass spectrometry for the novice*. CRC Press, 2013.

[42] Basil J Greber, Daniel Boehringer, Alexander Leitner, Philipp Bieri, Felix Voigts-Hoffmann, Jan P Erzberger, Marc Leibundgut, Ruedi Aebersold, and Nenad Ban. Architecture of the large subunit of the mammalian mitochondrial ribosome. *Nature* 505.7484 (2014), 515–519.

[43] Christoph Hage, Claudio Iacobucci, Anne Rehkamp, Christian Arlt, and Andrea Sinz. The First Zero-Length Mass Spectrometry-Cleavable Cross-Linker for Protein Structure Analysis. *Angewandte Chemie International Edition* 56.46 (2017), 14551–14555.

[44] Dominic Helm, Johannes PC Vissers, Christopher J Hughes, Hannes Hahne, Benjamin Ruprecht, Fiona Pachl, Arkadiusz Grzyb, Keith Richardson, Jason Wildgoose, Stefan K Maier, et al. Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics. *Molecular & Cellular Proteomics* 13.12 (2014), 3709–3715.

[45] Cherokee S Hoaglund-Hyzer, Young Jin Lee, Anne E Counterman, and David E Clemmer. Coupling ion mobility separations, collisional activation techniques, and multiple stages of MS for analysis of complex peptide mixtures. *Analytical chemistry* 74.5 (2002), 992–1006.

[46] Andrew N Holding. XL-MS: Protein cross-linking coupled with mass spectrometry. *Methods* 89 (2015), 54–63.

[47] Claudio Iacobucci and Andrea Sinz. To Be or Not to Be? Five Guidelines to Avoid Misassignments in Cross-Linking/Mass Spectrometry. *Analytical Chemistry* 89.15 (2017), 7832–7835.

[48] Amadeu H Iglesias, Luiz FA Santos, and Fabio C Gozzo. Collision-induced dissociation of Lys–Lys intramolecular crosslinked peptides. *Journal of the American Society for Mass Spectrometry* 20.4 (2009), 557–566.

[49] K.R. Jennings. Collision-induced decompositions of aromatic molecular ions. *International Journal of Mass Spectrometry and Ion Physics* 1.3 (1968), 227 –235. ISSN: 0020-7381.

[50] Susan Jones and Janet M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences* 93.1 (1996), 13–20.

[51] Stefan Kalkhof and Andrea Sinz. Chances and pitfalls of chemical cross-linking with amine-reactive N-hydroxysuccinimide esters. *Analytical and bioanalytical chemistry* 392.1-2 (2008), 305–312.

[52] Stefan Kalkhof and Andrea Sinz. Chances and pitfalls of chemical cross-linking with amine-reactive N-hydroxysuccinimide esters. *Analytical and bioanalytical chemistry* 392.1-2 (2008), 305–312.

[53] Abu B Kanu, Prabha Dwivedi, Maggie Tam, Laura Matz, and Herbert H Hill. Ion mobility–mass spectrometry. *Journal of Mass Spectrometry* 43.1 (2008), 1–22.

[54] Athit Kao, Chi-li Chiu, Danielle Vellucci, Yingying Yang, Vishal Rajesh Patel, Shenheng Guan, Arlo Randall, Pierre Baldi, Scott D Rychnovsky, and Lan Huang. Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Molecular & Cellular Proteomics* (2010), mcp–M110.

[55] Zachary Keltner, Jennifer A Meyer, Erin M Johnson, Amanda M Palumbo, Dana M Spence, and Gavin E Reid. Mass spectrometric characterization and activity of zinc-activated proinsulin C-peptide and C-peptide mutants. *Analyst* 135.2 (2010), 278–288.

[56] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24.21 (2008), 2534–2536.

[57] Lars Kolbowski, Marta L Mendes, and Juri Rappsilber. Optimizing the parameters governing the fragmentation of cross-linked peptides in a tribrid mass spectrometer. *Analytical Chemistry* 89.10 (2017), 5311–5318.

[58] Matthew A Lauber and James P Reilly. Structural analysis of a prokaryotic ribosome using a novel amidinating cross-linker and mass spectrometry. *Journal of proteome research* 10.8 (2011), 3604–3616.

[59] Alexander Leitner, Thomas Walzthoeni, and Ruedi Aebersold. Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using

LC-MS/MS and the xQuest/xProphet software pipeline. *Nature protocols* 9.1 (2014), 120–137.

[60] Alexander Leitner, Thomas Walzthoeni, and Ruedi Aebersold. Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline. *Nature protocols* 9.1 (2014), 120–137.

[61] Alexander Leitner, Roland Reischl, Thomas Walzthoeni, Franz Herzog, Stefan Bohn, Friedrich Förster, and Ruedi Aebersold. Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography. *Molecular & Cellular Proteomics* 11.3 (2012), M111–014126.

[62] Alexander Leitner, Thomas Walzthoeni, Abdullah Kahraman, Franz Herzog, Oliver Rinner, Martin Beck, and Ruedi Aebersold. Probing Native Protein Structures by Chemical Cross-linking, Mass Spectrometry, and Bioinformatics. *Molecular & Cellular Proteomics: MCP* 9.8 (2010), 1634.

[63] Diogo B Lima, Tatiani B de Lima, Tiago S Balbuena, Ana Gisele C Neves-Ferreira, Valmir C Barbosa, Fábio C Gozzo, and Paulo C Carvalho. SIM-XL: A powerful and user-friendly tool for peptide cross-linking analysis. *Journal of proteomics* 129 (2015), 51–55.

[64] Philip Lössl, Michiel van de Waterbeemd, and Albert JR Heck. The diverse and expanding role of mass spectrometry in structural and molecular biology. *The EMBO Journal* (2016), e201694818.

[65] John N Louris, R Graham Cooks, John EP Syka, Paul E Kelley, George C Stafford, and John FJ Todd. Instrumentation, applications, and energy deposition in quadrupole ion-trap tandem mass spectrometry. *Analytical Chemistry* 59.13 (1987), 1677–1685.

[66] Michael J MacCoss, Christine C Wu, and John R Yates. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Analytical chemistry* 74.21 (2002), 5593–5599.

[67] Alexander Makarov. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical chemistry* 72.6 (2000), 1156–1162.

[68] Alexander Makarov, Eduard Denisov, Alexander Kholomeev, Wilko Balschun, Oliver Lange, Kerstin Strupat, and Stevan Horning. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical chemistry* 78.7 (2006), 2113–2120.

[69] BA Mamyrin, VI Karataev, DV Shmikk, and VA Zagulin. The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution. *Soviet Journal of Experimental and Theoretical Physics* 37 (1973), 45.

[70] Raymond E March and Richard J Hughes. Quadrupole storage mass spectrometry (1989).

[71] A. D. McNaught, A. Wilkinson, M. Nic, J. Jirat, B. Kosata, and A. Jenkins. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. http://goldbook.iupac.org. Blackwell Scientific Publications, Oxford, 1997.

[72] Izhak Michaelevski, Noam Kirshenbaum, and Michal Sharon. T-wave ion mobility-mass spectrometry: basic experimental procedures for protein complex analysis. *Journal of visualized experiments: JoVE* 41 (2010).

[73] Wolfgang Mühlbacher, Sarah Sainsbury, Matthias Hemann, Merle Hantsche, Simon Neyer, Franz Herzog, and Patrick Cramer. Conserved architecture of the core RNA polymerase II initiation complex. *Nature communications* 5 (2014).

[74] DR Müller, P Schindler, H Towbin, U Wirth, H Voshol, S Hoving, and MO Steinmetz. Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis. *Analytical chemistry* 73.9 (2001), 1927–1934.

[75] Shuai Niu, Jessica N Rabuck, and Brandon T Ruotolo. Ion mobility-mass spectrometry of intact protein–ligand complexes for pharmaceutical drug discovery and development. *Current opinion in chemical biology* 17.5 (2013), 809–817.

[76] Jesper V Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy C-trap dissociation for peptide modification analysis. *Nature methods* 4.9 (2007), 709.

[77] Alexandre Panchaud, Pragya Singh, Scott A Shaffer, and David R Goodlett. xComb: A cross-linked peptide database approach to protein- protein interaction analysis. *Journal of proteome research* 9.5 (2010), 2508–2515.

[78] Siew Siew Pang, Richard Berry, Zhenjun Chen, Lars Kjer-Nielsen, Matthew A Perugini, Glenn F King, Christina Wang, Sock Hui Chew, Nicole L La Gruta, Neal K Williams, et al. The structural basis for autonomous dimerization of the pre-T-cell antigen receptor. *Nature* 467.7317 (2010), 844.

[79] Neha Patel, Florian Stengel, Ruedi Aebersold, and Matthew G Gold. Molecular basis of AKAP79 regulation by calmodulin. *Nature communications* 8.1 (2017), 1681.

[80] Vibhuti J Patel, Konstantinos Thalassinos, Susan E Slade, Joanne B Connolly, Andrew Crombie, J Colin Murrell, and James H Scrivens. A comparison of labeling and label-free mass spectrometry-based proteomics approaches. *Journal of proteome research* 8.7 (2009), 3752–3759.

[81] Richard H Perry, R Graham Cooks, and Robert J Noll. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass spectrometry reviews* 27.6 (2008), 661–699.

[82] T Prentice. Native, Ion Mobility and Crosslinking Mass Spectrometry: an Integrative approach Exploring Protein Stability and Flexibility. University College London, 2018.

[83] Steven D Pringle, Kevin Giles, Jason L Wildgoose, Jonathan P Williams, Susan E Slade, Konstantinos Thalassinos, Robert H Bateman, Michael T Bowers, and James H Scrivens. An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *International Journal of Mass Spectrometry* 261.1 (2007), 1–12.

[84] Juri Rappsilber. The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *Journal of structural biology* 173.3 (2011), 530–540.

[85] Thomson Reuters. Web of Science. (2012).

[86] H. E. Revercomb and E. A. Mason. Theory of plasma chromatography/gaseous electrophoresis. Review. *Analytical Chemistry* 47.7 (1975), 970–983.

[87] Oliver Rinner, Jan Seebacher, Thomas Walzthoeni, Lukas Mueller, Martin Beck, Alexander Schmidt, Markus Mueller, and Ruedi Aebersold. Identification of cross-linked peptides from large sequence databases. *Nature methods* 5.4 (2008), 315–318.

[88] P Roepstorff and J Fohlman. Letter to the editors. *Biological Mass Spectrometry* 11.11 (1984), 601–601.

[89] G. van Rossum and F.L. Drake. *Python Reference Manual, PythonLabs, Virginia, USA*. 2001. URL: `http://www.python.org` (visited on 02/14/2014).

[90] Emilio Sacristán and Andro A Solis. A swept-field aspiration condenser as an ion-mobility spectrometer. *IEEE Transactions on Instrumentation and Measurement* 47.3 (1998), 769–775.

[91] Abraham. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36.8 (1964), 1627–1639.

[92] Rico Schmidt and Andrea Sinz. Improved single-step enrichment methods of cross-linked products for protein structure analysis and protein interaction mapping. *Analytical and bioanalytical chemistry* 409.9 (2017), 2393–2400.

[93] Pragya Singh, Alexandre Panchaud, and David R Goodlett. Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique. *Analytical chemistry* 82.7 (2010), 2636–2642.

[94] Andrea Sinz. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *Journal of mass spectrometry* 38.12 (2003), 1225–1237.

[95] Andro A Solis and Emilio Sacristán. Designing the measurement cell of a swept-field differential aspiration condenser. *Revista mexicana de física* 52.4 (2006), 322–328.

[96] Catherine A Srebalus Barnes, Amy E Hilderbrand, Stephen J Valentine, and David E Clemmer. Resolving isomeric peptide mixtures: a combined HPLC/ion mobility-TOFMS analysis of a 4000-component combinatorial library. *Analytical chemistry* 74.1 (2002), 26–36.

[97] Florian Stengel, Ruedi Aebersold, and Carol V Robinson. Joining forces: integrating proteomics and cross-linking with the mass spectrometry of intact complexes. *Molecular & Cellular Proteomics* 11.3 (2012), R111–014027.

[98] DP Stevenson. Mass Spectrometry and its Applications to Organic Chemistry. *Journal of the American Chemical Society* 83.12 (1961), 2787–2787.

[99] Jingchuan Sun, Yi Shi, Roxana E Georgescu, Zuanning Yuan, Brian T Chait, Huilin Li, and Michael E O'donnell. The architecture of a eukaryotic replisome. *Nature structural & molecular biology* 22.12 (2015), 976.

[100] Dan Tan, Qiang Li, Mei-Jun Zhang, Chao Liu, Chengying Ma, Pan Zhang, Yue-He Ding, Sheng-Bo Fan, Li Tao, Bing Yang, et al. Trifunctional cross-linker for mapping protein-protein interaction networks and comparing protein conformational states. *Elife* 5 (2016), e12509.

[101] Xiaoting Tang and James E Bruce. A new cross-linking strategy: protein interaction reporter (PIR) technology for protein–protein interaction studies. *Molecular biosystems* 6.6 (2010), 939–947.

[102] Xiaoting Tang, Gerhard R Munske, William F Siems, and James E Bruce. Mass spectrometry identifiable cross-linking strategy for studying protein- protein interactions. *Analytical chemistry* 77.1 (2005), 311–318.

[103] JA Taraszka, AE Counterman, and DE Clemmer. Gas-phase separations of complex tryptic peptide mixtures. *Fresenius' journal of analytical chemistry* 369.3-4 (2001), 234–245.

[104] Konstantinos Thalassinos, Megan Grabenauer, Susan E Slade, Gillian R Hilton, Michael T Bowers, and James H Scrivens. Characterization of phosphorylated peptides using

traveling wave-based and drift cell ion mobility mass spectrometry. *Analytical chemistry* 81.1 (2008), 248–254.

[105] Konstantinos Thalassinos, Arun Prasad Pandurangan, Min Xu, Frank Alber, and Maya Topf. Conformational states of macromolecular assemblies explored by integrative structure calculation. *Structure* 21.9 (2013), 1500–1508.

[106] Thermo Fisher Scientific. *Thermo Velos Pro Schematic*. 2017. URL: `http://planetorbitrap.com/orbitrap-elite#tab:schematic` (visited on 05/17/2017).

[107] Thermo. *Exactive Plus Software Manual*. Thermo Fisher Scientific Inc., 2012, pp. 3–34.

[108] Thermo ThermoScientific. Normalized Collision Energy Technology. *Product Support Bulletin* 104 (2009).

[109] Karsten Thierbach, Alexander von Appen, Matthias Thoms, Martin Beck, Dirk Flemming, and Ed Hurt. Protein interfaces of the conserved Nup84 complex from Chaetomium thermophilum shown by crosslinking mass spectrometry and electron microscopy. *Structure* 21.9 (2013), 1672–1682.

[110] Joseph J Thomson. XIX. Further experiments on positive rays. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 24.140 (1912), 209–253.

[111] Michael J Trnka, Peter R Baker, Philip JJ Robinson, AL Burlingame, and Robert J Chalkley. Matching cross-linked peptide spectra: only as good as the worse identification. *Molecular & Cellular Proteomics* 13.2 (2014), 420–434.

[112] W.R. Rays of Positive Electricity and their Application to Chemical Analysis. *Nature* 92.1914/01/15/online (1914), 549–550.

[113] Alistair Wallace. A High-Resolution Ion Mobility: Mass Spectrometry Platform for Breakthrough Discoveries in Life Science Research and the Pharmaceutical Industry. *American laboratory* 42.6 (2010), 13–17.

[114] Thomas Walzthoeni, Manfred Claassen, Alexander Leitner, Franz Herzog, Stefan Bohn, Friedrich Förster, Martin Beck, and Ruedi Aebersold. False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nature methods* 9.9 (2012), 901–903.

[115] Sebastian Wiese, Kai A Reidegeld, Helmut E Meyer, and Bettina Warscheid. Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics* 7.3 (2007), 340–350.

[116] Bing Yang, Yan-Jie Wu, Ming Zhu, Sheng-Bo Fan, Jinzhong Lin, Kun Zhang, Shuang Li, Hao Chi, Yu-Xin Li, Hai-Feng Chen, et al. Identification of cross-linked peptides from complex samples. *Nature methods* 9.9 (2012), 904–906.

[117] Haizhen Zhang, Xiaoting Tang, Gerhard R Munske, Nikola Tolic, Gordon A Anderson, and James E Bruce. Identification of protein-protein interactions and topologies in living cells with chemical cross-linking and mass spectrometry. *Molecular & Cellular Proteomics* 8.3 (2009), 409–420.

[118] Roman A Zubarev and Alexander Makarov. Orbitrap mass spectrometry. 2013.